VOLUME 62
NUMBER 8

WHOLE NO. 295
1948

Psychological Monographs:
General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology* with the *Psychological Monographs*

HERBERT S. CONRAD, *Editor*

Characteristics and Uses of Item-Analysis Data

By

HERBERT S. CONRAD

formerly of the
Educational Testing Service
Princeton, N.J.

Accepted for publication, June 15, 1948

Price $1.00

Published by
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
Publications Office
1515 MASSACHUSETTS AVE. N.W., WASHINGTON 5, D.C.

FOREWORD

DURING the war, the College Entrance Examination Board contracted to carry out for the U. S. Navy, through Project N-106 (under the Applied Psychology Panel of the Office of Scientific Research and Development) various developmental and research programs in the field of aptitude and achievement testing. A characteristic methodological feature of this work was the application of item analysis in the evaluation of each item of each test. The method of item analysis adopted by Project N-106 was (with occasional exceptions) the method which had been adopted as "standard" by the College Entrance Examination Board for its own tests. (This is also the method generally applied by the Educational Testing Service, which now provides technical service to the College Entrance Examination Board.) While the method cannot be fairly described in a phrase, a brief characterization would mention that the method is based essentially on use of the biserial correlation coefficient, and of the mean criterion scores of those who choose, respectively, alternatives 1, 2, 3, . . . n (or no alternative at all) of an n-choice multiple-choice item. Practically speaking, this method requires the use of modern electric tabulating equipment—it would be too time consuming by manual techniques.

Because Project N-106 was making very extensive use of item analysis by the method indicated, and because no systematic, detailed consideration of the method had appeared, the author was asked, as a member of Project N-106, to prepare the present report. The report was made immediately available to the Navy, and routinely classified as "restricted." With the declassification of documents after the war, permission was obtained to print the report, but with the routine proviso that the report be printed without change. The latter proviso or requirement explains several shortcomings of the report—such as the failure to give recognition to alternative systems of item analysis. A second criticism applies to the title of the report. In the context in which the report was originally issued, the title appeared quite acceptable; in the present context, however, the title is obviously too broad—again because alternative systems of item analysis are not considered. In mitigation, let us suggest that much of what is said here regarding the uses of item analysis applies equally to other systems. Further, it is hoped that the content of the report will be considered more important than the title.

For unusually careful and helpful reading of the original manuscript, and for encouragement to publish the report, the writer is indebted to Professor J. M. Stalnaker (Contractor's Technical Representative for Project N-106), to Dr. H. O. Gulliksen (Project Director), and to Dr. N. O. Frederiksen and Mr. Donald A. Peterson (colleagues in Project N-106). For additional encouragement toward publication, the writer wishes to express thanks to Dr. Walter S. Hunter and Dr. Charles W. Bray (successively chairmen of the Applied Psychology Panel), to Dr. Dael Wolfle (Panel consultant), and to Mr. Henry Chauncey (President of the Educational Testing Service). Special acknowledgment is due to the Educational Testing Service for providing funds to defray publication costs.

HERBERT S. CONRAD

TABLE OF CONTENTS

FOREWORD iii

I. INTRODUCTION 1

II. TYPES OF INFORMATION SUPPLIED BY ITEM ANALYSIS 1

III. INFORMATION CONCERNING THE SAMPLE ATTEMPTING EACH ITEM 3

A. Number of Individuals Attempting Each Item (N_t) 3

B. Mean (M_t) and Standard Deviation (σ_t) of Those Attempting Each Item 5

IV. INFORMATION CONCERNING THE ITEM-AS-A-WHOLE 7

A. Ease of Each Item (p) 7

B. Difficulty of Each Item in Terms of "Δ" 9

C. Comparison between Δ and p 11

D. Biserial Correlation ($r_{bis.}$) between Item and Criterion 12

1. Some Statistical Aspects of Biserial r 13

2. Effect of Use of N_t vs. Base N in Formula for Biserial r 14

3. Meaning of Biserial r in Terms of "Internal Consistency" and Item-Validity 15

4. Choice of Test-Criterion 16

5. Factors Affecting the Interpretation of Biserial r 17

a. Biserial r in Relation to Percentage of Successful Attempts (p) 17

b. The "Probable Error" (PE) or Sampling Fluctuation of Biserial r 18

c. Biserial r in Relation to Variability or "Range of Talent" of the Group Attempting the Item (σ_t) 20

d. Biserial r in Relation to Speed 21

e. Biserial r in Relation to Length of Subtest 22

f. Biserial r in Relation to Reliability of the Criterion 23

g. Limitations of Biserial r as a Measure of Item-Validity 23

V. INFORMATION CONCERNING THE ALTERNATIVES WITHIN EACH ITEM 26

VI. NEED FOR INTERPRETATION 27

VII. USES OF ITEM-ANALYSIS DATA 28

A. Provision of Objective, Quantitative Evidence Concerning Individual Items 28

B. Improvement of Test Items 28

C. Item Analysis vs. Expert Judgment in the Elimination of Inferior Items 30

D. Improvement of Distribution of Item-Difficulty 32
E. Improvement of Reliability 32
F. Improvement of Independence of a Test or Subtest 34
G. Improvement of Correlation between Subtest and External Criterion 34
H. Stimulation of Hypotheses and Insights 35

VIII. Recommendations 36
A. Utilization of Item-Analysis Results 36
B. Verification of Subjective Judgments concerning Items 36
C. Elimination of Effect of Speed upon Functional Homogeneity of Items 36
D. Time-Limits and Make-up of Experimental Tests 37
E. Size of Sample 38
F. Restriction of Item Analysis to Experimental Forms 39
G. Discrimination in the Calculation of $r_{bis.}$ 39
H. Determining the Reliability of the Experimental Form 40
I. Correlation with an External Criterion 41

IX. Summary 42
Appendix 48

CHARACTERISTICS AND USES OF ITEM-ANALYSIS DATA

I. INTRODUCTION

PROJECT N-106 has made item analyses of many of the tests employed in the selection program of the Navy. The purpose of the present Report is primarily to supply a general, explanatory appraisal of the information yielded by the particular type of item analysis performed by this Project.

II. TYPES OF INFORMATION SUPPLIED BY ITEM ANALYSIS

ALL the information furnished directly by item analysis is objective and quantitative. The information supplied in the item analyses of this Project may be classified into three main categories and nine sub-categories, as follows:

1. Information concerning the item-as-a-whole (as distinguished from the *individual* choices or alternatives offered by the item). The information includes:

a. A measure of the *ease* of the item. This is the percentage of successful attempts to answer the item; it is designated by the symbol p, and calculated by the formula, $p = 100\ (N_c/N_t)$, where N_c represents the number of correct responses, and N_t the number of attempts to answer the item. For a more complete definition of N_t, see section 3,a below. The higher the value of p, the easier the item.

In the reports of this project, p is sometimes written as a proportion instead of as a percentage; e.g., .75, instead of 75 per cent. The two modes of expression are, of course, equivalent.

b. A measure of the *difficulty* of the item. This measure, designated by the Greek letter "Δ" (delta) is computed in a manner quite different from p, and expressed in terms of a different unit. The definition and explanation of Δ may best be reserved till section IV,B below.

c. A measure of the *correlation* between the item and some criterion. Usually, the criterion is the score on the *subtest* of which the item is a part; if the test is not divided into subtests, the score on the *total* test is employed. Occasionally, an external criterion (such as grades in service school, or grades in a specific subject in service school) may be employed. In the present Report, we shall use the term "*item-criterion correlation*" to mean the *biserial correlation* between the item and whatever particular criterion has been employed in the given case—whether total-test score, subtest score, or school grade. The conventional symbol for biserial correlation is "$r_{bis.}$"

d. The number of individuals who "skipped" the item. A person is judged to have skipped an item if he failed to record a response to the item, yet answered one or more subsequent items in the subtest (or in the total test, if the test is not divided into subtests). Normally, the number of cases skipping an item is small, since the directions for tests in use in the Navy generally prescribe that men make the best choice that they can, rather than to leave an item entirely unmarked. A skipped item is counted as an unsuccessful attempt to answer the item, and hence is included in N_t (see section 3,a below).

"Skipped" items are distinguished from "omitted" items. An item is counted as

omitted, if no answer is recorded for that item or any subsequent item in the subtest. (In the case of the last item of a subtest, this is counted as omitted simply if no answer is recorded for it.) No score-credit is given for either skipped or omitted items.

2. Information concerning the individual choices or alternatives offered by the item. Such information includes:

a. The number of individuals (among those attempting the item) who selected a given alternative in the item as the answer; this number is designated by the symbol, n.

b. The mean criterion-score of those selecting a given alternative in the item (as well as the mean criterion-score of those who "skipped" the item). The symbol for mean criterion-score is M. In calculating M, use is made of a transformation of the raw criterion-scores; this transformation is such that (within errors of rounding and grouping) the mean of the transformed scores of the total sample is 13.000 and the standard deviation is 4.000. The correlation between the original and the transformed criterion-scores is 1.00. The purposes of the transformation are first, to provide a standard or uniform scale of criterion-scores,[1] and second, to facilitate the use of mechanical-tabulation equipment.

3. Information concerning the sample attempting to answer each item. This information includes the following:

a. N_t, the number of persons who attempted (or tried) to answer each item. N_t cannot exceed "Base N," the total number of persons measured on the criterion. An individual is considered to have "attempted" an item if he has recorded an answer either to this item or to any *subsequent* item in the subtest[2] of which the item is a part. The assumption underlying this definition of N_t is that the recruit works systematically through each subtest, mentally attempting all items up to and including the last one for which an answer is recorded. From the definition, it follows that N_t may decrease from an earlier to a later item of a subtest, but cannot increase—since a person attempting the later item is counted as having attempted the earlier. Alternative definitions of N_t, equivalent to that given above, are, first, N_t = No. recording an answer to the item + no. "skipping" the item; and, second, N_t = Base N − no. omitting the item. Except in special instances or for special purposes, *all the item-data reported by this project are based on the sample defined by* N_t.

b. M_t, the mean (transformed) criterion-score of those who attempted to answer the item.

c. σ_t, the standard deviation of the (transformed) criterion-scores of those who attempted the item.

It may be added that the various measures described above—p, Δ, $r_{bis.}$, n, M, N_t, M_t, and σ_t—are identical with those employed for years by the College Entrance Examination Board, under whose jurisdiction this Project has operated. Many other systems or techniques of item analysis are, of course, available; it is not our purpose here to enter into a lengthy discussion or comparison of the different systems that could be adopted. Suffice it to say that, under the Project's operating conditions, the measures defined above offered the greatest promise of facility and accuracy

[1] A prime pre-requisite for such a standard scale of criterion-scores is that the sample of cases should itself be standard or uniform from one test to the other. A second requirement is that the distribution of test scores should be standard or uniform: fully comparable scores cannot be obtained for tests which do not have comparable distributions.

[2] If the test is not divided into subtests, the word "test" should be substituted for "subtest" in this definition.

of computation; they are also measures which have proved useful in the lengthy and extensive experience of the College Entrance Examination Board.

In the remainder of this Report, the characteristics of each of the various measures (p, Δ, $r_{bis.}$, n, M, etc.) will be considered in some detail. A specimen "item-analysis sheet" is given in the Appendix.

III. INFORMATION CONCERNING THE SAMPLE ATTEMPTING EACH ITEM

A. Number of Individuals Attempting Each Item (N_t)

In general, the more closely the value of N_t for each item approaches Base N (the total number of cases measured on the criterion), the better. Each item, presumably, has some merit; and it is obviously desirable that this merit should be applied directly to a large number of individuals, rather than to only a small proportion of Base N. The only exception to these remarks occurs in the case of a test intended to measure mainly speed of performance.

A second desideratum of N_t is that it be numerically large. All of the statistical measures for an item (M_t, σ_t, p, etc.) are, of course, subject to sampling error; the larger the value of N_t, the smaller this error is likely to be. It is important to remember that the value of N_t, even if satisfactorily large for the first item of a subtest, may have diminished considerably for items nearer to the end of the test. Table 1 illustrates this point. In this table are given values of N_t for the first, middle, and last items of all the tests or subtests of the Navy Basic Classification Test Battery (Form I); the figures in Table 1 are based on a national sample of 500 cases drawn from six naval training stations. Table 1 shows how seriously fallacious it may be to think in terms of Base N, or the value of N_t for the first item, when items in the later portion of a subtest are under consideration.

The need for a numerically large value of N_t applies especially with reference to $r_{bis.}$ (the biserial correlation between item and criterion), because of the inherently high sampling error ("*PE*") of biserial r. It also applies with special force in connection with n (the frequency with which each alternative within an item is chosen), and M (the mean score of the individuals choosing each alternative). Unless N_t for an item is large, the values of n for the various response-alternatives within an item must on the average be small, thus rendering differences between the n's of very questionable reliability. Similarly, the values of M for each alternative within an item will be based on a small number of cases, again rendering differences unreliable.

If N_t is considerably smaller than Base N (say only half as large), it might be supposed that the sample represented by N_t is rather strongly selected—since, presumably, it is mainly the less capable individuals who tend to drop out. A much more direct and generally dependable measure of selection, however, is provided by M_t and σ_t (the mean and standard deviation, respectively, of those attempting to answer the item). To illustrate this point, Table 2 presents the values of N_t, M_t, and σ_t for selected items from two tests;[1] for each of these tests, the value of Base N equals 500

[1] The "Surface Development" test listed in Table 2 is a subtest of the Mechanical Aptitude Test.

TABLE 1

VALUES OF N_t FOR THE FIRST, MIDDLE, AND LAST ITEMS OF VARIOUS TESTS OR SUBTESTS

Item	Values of N_t for				
	Sentence Completion	Opposites	Analogies	Reading	Arithmetical Reasoning
First item	500	500	500	500	500
Middle item	491	494	499	498	493
Last item	208	364	443	310	220
	Block Counting	Mechanical Comprehension	Surface Development	Tool Relationships	Mechanical Information
First item	499	500	489	500	500
Middle item	429	498	378	498	500
Last item	106	338	95	421	296

(constituting a national sample, drawn from six naval training stations). The entries in each line of the table are matched, as closely as the data permit, with respect to N_t. After the first item, it will be observed that groups virtually identical with respect to N_t may differ considerably with respect to M_t and σ_t. Thus, when $N_t = 347$ for the Reading Test, $M_t = 13.6$ and $\sigma_t = 4.1$; whereas for the same value of N_t in the Surface Development Test, M_t is considerably higher, 14.9, and σ_t is considerably lower, 3.0.

Unlike other measures to be considered later, N_t does not represent an inherent property of an item. N_t depends primarily on the position of the item in the subtest, and the time-limit set for the subtest. Another factor is the rate of increase in difficulty from earlier to later items of the subtest: in a steeply-graded power test, there is a definite tendency for men to stop recording answers, after they have reached a point which is obviously beyond their ability. (This assumes, of course, that the items of the subtest are arranged more or less in order of difficulty.)

The value of M_t may be spuriously low if, after reaching a certain point in the subtest, the subject mentally attempts additional items but fails to record any answer for these items. On the other hand, the value of N_t may be somewhat too large if, after encountering a few items that are beyond his ability, the subject guesses at the answers of the remaining items with only perfunctory effort—such "attempts" being only half-hearted at best and definitely different from the attempts in the early and easier part of the test. Finally, the value of N_t will be too large for certain items if, after answering (say) 15 consecutive items, the subject "takes a crack" at the last item or so of the test, without attempting to answer the intervening items. According to the definition of N_t, the intervening items are counted as having been attempted, because a *subsequent* item has been attempted. Such erratic responses, however, occur only seldom, provided that (*a*) the items in each subtest are arranged in order of difficulty, (*b*) the test-directions are adequate in their emphasis on a systematic approach, and (*c*) the proctoring is competent.—Under typically good conditions, the sources of error mentioned in this paragraph will generally be of only minor importance. Or-

dinarily, it seems safe to accept the value of N_t as a fairly accurate expression of what it is intended to measure.

The discussion above has assumed that Base N (of which N_t is a subsample) is a fair and representative sample of the population which the test aims to measure. The fulfillment of this condition is obviously of paramount importance.

B. Mean (M_t) and Standard Deviation (σ_t) of Those Attempting Each Item

The nature of the sample attempting each item is indicated directly by two measures: M_t, the mean of the transformed criterion-scores of those attempting the item; and σ_t, the standard deviation of the transformed criterion-scores of those attempting the item. As previously stated, the mean transformed criterion-score of the total group (Base N) is 13.0; the standard deviation for the total group is 4.0. (Errors due to grouping or rounding may, of course, cause slight departure from these norms.) An example of the changes in M_t and σ_t which may be expected as one proceeds from early to later items of a subtest is given in Table 2. In this table, there is little uniformity in the rise of M_t from early to later items, or in the decline of σ_t—though the rise in M_t occurs more regularly than the decline in σ_t. The factors which, in general, determine the trend in M_t and σ_t are:

1. The time-limit for the test: the more sharply limited the time, the steeper the rise in M_t and the drop in σ_t. With a short time-limit, the group attempting the later items tends to be relatively homogeneous and superior—partly because of superior speed of performance, and partly because of the positive correlation which usually prevails between speed and level of ability.

2. The rate of increase in difficulty from early to later items in the subtest: the more rapid the increase, the greater the change in M_t and σ_t. This factor would not operate if all persons attempted each item. As already mentioned, however, there is a tendency for

TABLE 2

Values of N_t, M_t, and σ_t, for Selected Items from Two Tests

TEST							
Reading				Surface Development			
Item No.	N_t	M_t	σ_t	Item No.	N_t	M_t	σ_t
1	500	13.0	4.0	1	489*	13.2	3.9
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
23	449	13.3	4.0	9	451	13.8	3.5
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
26	393	13.5	4.0	18	397	14.4	3.1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
28	347	13.6	4.1	21	347	14.9	3.0

* Base $N = 500$, but 11 cases failed to answer any of the items in the Surface Development subtest of the Mechanical Aptitude Test.

individuals taking a test to stop recording answers, after they have reached a point which is obviously beyond their ability. In such an event, the persons attempting the later items tend to be more selected and homogeneous than those attempting the early items.

3. The correlation between number of items attempted (or speed of performance) and level of ability: the higher the correlation, the greater the change in M_t and σ_t. A high positive correlation between speed and ability-level reinforces the effects already noted in *1* and *2* above.

4. The homogeneity or internal consistency of the items in the subtest: the higher the homogeneity, the greater the changes in M_t *and* σ_t. If the items in a subtest are lacking in homogeneity, this tends to dampen the selective effect of differences in speed of performance and level of ability—since a person who is exceptionally quick or capable on one type of item will not, in general, be equally quick or able on items of functionally different type. A measure of the homogeneity between an item and the remaining items of a subtest is provided by biserial r (see section IV,D, below).

In Table 2, the changes in M_t and σ_t are greater for scores in Surface Development than for scores in Reading. The statistics presented in Table 3 for these two tests are indicative of the role of factors *1*, *2*, and *4* just mentioned. It will be noticed, in Table 3, that Surface Development shows the higher median value of biserial r (factor 4); in addition, Surface Development shows the larger drop in N_t from first to last item—implying either a more highly restricted time-limit, or a steeper gradation in item-difficulty, or both (factors 1 and 2).

TABLE 3

SELECTED STATISTICS FOR READING AND SURFACE DEVELOPMENT TESTS

Test	N_t for First Item of Test	N_t for Last Item of Test	Median Biserial r
Reading	500	310	.52
Surface Development	489	95	.66

As for any mean, the sampling error of M_t may be estimated from the number of cases (N_t) and the standard deviation of the distribution (σ_t). On the assumption of a normal distribution, the same applies to the sampling error of σ_t.

M_t and σ_t are of interest not only as a description of the sample attempting each item, but also for their bearing on other measures. Both M_t and σ_t are required for the calculation of "Δ" (see section IV,B, below); M_t is also of importance in the interpretation of p (section IV,A), and σ_t is pertinent in the interpretation of r_{bis}. (section IV,D).

IV. INFORMATION CONCERNING THE ITEM-AS-A-WHOLE

THE TWO main measures considered in this section are p (a measure of the ease of the item) and $r_{bis.}$ (the item-criterion correlation). It is on the basis of p and $r_{bis.}$ usually, that an item is either retained or rejected for use in a test. A measure of item-difficulty, designated by the Greek letter "Δ" (delta) will also be considered. The only other information relating to the item-as-a-whole is the number of persons who "skip" the item. This number is usually quite small,[1] and hence does not require extended consideration. An excessively large number of "skips" may, however, occur if the item is very much more difficult than its neighbors; or if the test-directions have failed to emphasize the desirability of working consecutively from the first item of the subtest to each succeeding item.

A. EASE OF EACH ITEM (p)

The quantity p states the percentage of successful attempts to answer the item. Expressed as a formula,

$$p = 100\,(N_c/N_t)$$

where N_c is the number of individuals answering the item correctly, and N_t is the number of individuals attempting to answer the item. Thus, if 400 men attempted to answer an item, and 240 of these answered correctly, $p = 60$.

The formula for the degree of fluctuation to be expected in p as a result of random sampling is:

$$PE_p = .6745\sqrt{\frac{p(100-p)}{N_t}}.$$

This formula is applicable except when N_t is quite small (below 50), or when p is close to 0 or 100. Ideally, the value of p to be inserted in the formula is the "true" value of p (i.e., the "universe" value, or the value obtained when N_t is extremely large). Tolerably fair results are generally obtained, however, when the empirical value of p is substituted for the (unknown) "true" value. From the formula, it is evident that the more closely p approaches 50, the larger the sampling error of p. This is unfortunate, in view of the fact that most items in typical tests are selected to be of roughly medium difficulty (i.e., with p not very far from 50). Actually, however, the *PE* of p is never large enough to be an important practical issue, *so long as* N_t *is fairly large* (say 400 or more). Thus, in the case of our illustration of the preceding paragraph ($p = 60$, $N_t = 400$), the *PE* of the item is

$$.6745\sqrt{\frac{60(40)}{400}},$$

or only 1.65. Suppose, however, that N_t had been 100 instead of 400—as not infrequently happens for the later items of a subtest designed to measure both speed and power of performance. In such a case, the *PE* of our illustrative item would be twice as great, i.e., 3.30 instead of 1.65. Taking ±2PE as the minimum range of variation which must be given practical consideration, an item whose true p is 60 may (when N_t equals only 100) turn up empirically as any value between (60−6.6) and (60+6.6), or between 53.4 and 66.4. The difference between these limiting values is 13.2—which seems too large to be tolerated. In selecting a sample for the collection of item-analysis data, it must be remembered that the sampling errors of the statistics will depend not on Base N but on N_t; and for the later items of a subtest, N_t may be

[1] Because the number of "skips" is usually too small to warrant serious consideration, it is not customary to convert the number of "skips" into a proportion or percentage of N_t.

much smaller than Base N.

Since the quantity N_t appears as the denominator in the formula, $p = 100\ (N_c/N_t)$, a spuriously large or small value of N_t would result, respectively, in a spuriously low or high value of p. Normally, however, the value of N_t is accurate enough for practical purposes (see section III,A, above). A more important issue is whether N_t or Base N serves better as the denominator in the formula for p. This issue might be settled by taking the view that the use of N_t and Base N both provide useful information; to the extent that N_t differs from Base N, the information is different, but scarcely subject to a judgment of better or worse. In the discussion which follows, the assumption is that we wish to know the value of p that would be found if all members of the sample explicitly attempted to answer the item; and the question is whether this information is obtained better by the use of N_t or Base N in the formula for p.

For the early items of a subtest, the numerical values of Base N and N_t are likely to be either identical or closely alike; in such cases, the question whether N_t or Base N serves better as the denominator for p is of no practical importance. Consider, however, the following data for item no. 71 of a certain 80-item test. For this particular test, Base $N = 500$; for the item in question, $N_t = 400$. The number of correct answers to the item is 164. If Base N is employed as the denominator in computing p, $p = 100\ (164/500) = 33$; if N_t is employed, $p = 100\ (164/400) = 41$. The percentage, 33, is *too low* as an index of the ease of item no. 71, because some of those who failed to attempt this item would (either by knowledge or chance) have gotten the item right, had they attempted it. On the other hand, the p-value of 41, obtained by use of N_t, is *too high*, because the group represented by N_t is somewhat superior to the total sample—its M_t is 13.7 instead of 13.0. The chief source of this superiority is doubtless the correlation between speed and power: except in a pure-speed test, those answering the later items are generally not only faster, but also more likely to answer correctly items at a higher level of difficulty.

Base N is the proper denominator to use in the formula for p, if it is assumed that a perfect positive relation exists between speed and power; or, more specifically, if it is assumed that a person who failed to reach an item would have failed to answer the item correctly. The assumption of so close a correlation between speed and item-score is unreasonable, since merely by chance a certain proportion of the answers to a multiple-choice item will be correct. The use of N_t in the denominator of the formula for p involves the assumption that speed and power are completely uncorrelated. Here it is assumed that, had more time been allowed, those who failed to reach an item would perform the same as those who did reach the item. Recent studies in the experimental literature favor the view that the relation between speed and power, while positive, is rather low. From these studies, it would appear that neither the assumptions underlying the use of N_t or Base N are fully justified; but the assumption underlying N_t seems better supported than that underlying Base N.

A complicating factor which deserves some attention relates to the arrangement of items in a test. Ordinarily, the items of each subtest are arranged in order of difficulty for the *average* individual. This is not, of course, the same as

the order of difficulty for each *individual tested*. A recruit may strike a region of a test which, for him, proves excessively difficult; in this event, he may easily spend an excessive amount of time on the (for him) difficult section, and even be discouraged from attempting items beyond it—on the theory that whatever comes later in the test is probably still harder and quite beyond his ability. In this way, through lack of time or lack of encouragement, the recruit may fail to attempt later items which (in his particular case) may be easier than the ones he has failed on. We do not know to what extent this occurs; if it does occur, the use of N_t makes better allowance for this factor than does Base N.

In the case of a test measuring mainly speed of performance, the use of Base N would result in p-values which reflect the *position* of an item in the subtest, far more than inherent difficulty. For the items of such a test the use of N_t is preferable. On the other hand, in a power test with unlimited time, Base N is preferable; but in such a test, N_t should be equal or closely similar to Base N, so that the practical advantage of Base N over N_t would typically be slight.

On the whole, the use of N_t seems preferable to Base N for determining the ease or difficulty of a test-item. As previously indicated, the use of N_t tends in general, to make p too high (the item appears easier than it really is); while the use of Base N tends in general, to make p too low. The greater the difference between N_t and Base N, the greater the likelihood of error in p. Ordinarily, it will hardly be worthwhile to compute two p's, one based on N_t, the other on Base N. According to our analysis, the use of N_t typically yields the more valid estimate.

B. Difficulty of Each Item in Terms of "Δ"

In some reports of this Project, use has been made of a measure of item-difficulty designated by the Greek letter "Δ" (delta). This measure was devised by C. R. Brolyer and C. C. Brigham. The Δ-value for an item is measured along the same scale as the transformed[2] criterion-scores of the group attempting the item. For each item, the percentage of cases (of those attempting the item) with scores above a certain transformed criterion-score equals p, the percentage of successful attempts to answer the item; this particular transformed criterion-score is the value of Δ for the particular item.[3] The higher the value of Δ, the more difficult the item.

The general formula for Δ is

$$\Delta = M_t + x'\sigma_t.$$

In this formula, x' is the abscissal value, in a unit normal curve, corresponding to the value of p (values of x' are negative when p exceeds 50, and positive when p falls below 50); M_t and σ_t are the mean and standard deviation, respectively, of the transformed criterion-scores of those attempting the item. The term x' involves the assumption that the distribution of criterion-scores of those attempting the item is normal. The multiplication of x' by σ_t serves to convert the unit of measurement from 1 (the σ of the unit normal curve) to the corresponding unit (σ_t) of the distribution of trans-

[2] As mentioned in section II of this Report, "transformed" criterion-scores are criterion-scores corrected to a standard distribution such that, for the total sample (Base N), the mean is 13.0 and the standard deviation is 4.0. When $N_t \neq$ Base N, the value of M_t generally exceeds 13.0 and the value of σ_t generally falls below 4.00. See section III,B.

[3] This is the definition given by C. R. Brolyer and C. C. Brigham. See *A Study of Error*, by C. C. Brigham (New York: College Entrance Examination Board, 1932), p. 356.

formed criterion-scores of those attempting the item. The term M_t takes account of the fact that the higher the mean score of those attempting the item, the greater the item-difficulty denoted by a given value of p. Following are two illustrations:

> Suppose that, for the sample attempting a given item, $M_t = 13.2$ and $\sigma_t = 3.9$; p, the percentage of successful attempts, equals (say) 84. Reference to a table of the normal curve shows that the value of x' corresponding to 84 is -1.00. Hence, the value of Δ for the item is $13.2 - 1.00(3.9) = 9.3$. As a second example, suppose that, for the sample attempting an item, $M_t = 14.0$ and $\sigma_t = 3.6$, and p again equals 84. Then $\Delta = 14.0 - 1.00(3.6) = 10.4$. In this second example, the value of Δ is higher than before (denoting greater item-difficulty), because it required a comparatively superior group (with $M_t = 14.0$ vs. 13.2) to achieve the same percentage of success ($p = 84$).

Because this Project has made comparatively little use of Δ, a detailed technical discussion of Δ will not be attempted in this place. Two observations may, however, be in order. First, if values of Δ for different tests are to be compared, it is essential that the samples to whom the tests are administered be comparable; otherwise, the scales of values of transformed criterion-scores, along which Δ is measured, will not be comparable. Second, the calculation of Δ fails to take account of the effect of guessing or chance-success (the same remark applies also to p). For this failure to correct for chance, several reasons may be offered: (1) The variation between corrected and uncorrected values is negligible, unless there are wide individual differences in the total number of items attempted by different individuals (and this seldom occurs when the time-limit for a test is generous). (2) The correction for chance is more likely to be important when comparisons are made between items for which the probability of chance-success is quite different—e.g., two-choice vs. five-choice items; but such comparisons in the work of this Project are uncommon. (3) The proper correction for chance is not entirely obvious; thus, it seems more likely that the failures on a very *hard* two-choice item are actually matched by an equal (or greater) number of chance-successes, than are the failures on a very *easy* item. Finally, as a practical consideration, (4) it is computationally simpler and more economical not to make any correction for guessing.

Table 4 below shows how variations in the difficulty of items from one test to another are reflected in Δ. (The figures in Table 4 are based on data from a national sample of 500 cases, tested on the Navy Basic Classification Test Battery, Form I.)

TABLE 4

MEANS AND STANDARD DEVIATIONS OF VALUES OF Δ FOR THE ITEMS IN EIGHT TESTS OF THE BASIC CLASSIFICATION TEST BATTERY, FORM I

Test or Subtest	Mean Value of Δ	S.D. of Values of Δ
Sentence Completion	11.9	3.8
Opposites	12.6	3.4
Analogies	13.1	2.8
Reading	12.8	2.9
Arithmetical Reasoning	12.5	3.1
Block Counting	12.2	2.3
Mechanical Comprehension	12.3	1.9
Surface Development	12.7	2.3

The range of mean values of Δ in Table 4 is from 11.9 (for Sentence Completion) to 13.1 (for Analogies). The range of S.D.'s of values of Δ is from 1.9 (for Mechanical Comprehension) to 3.8 (for Sentence Completion). According to these data, the items in different tests tend to be similar with respect to *average* difficulty; but the differences of difficulty among the individual items tend

to be much greater for some tests (e.g., Sentence Completion) than for others (e.g., Mechanical Comprehension).

A comparison between Δ and p is presented in the next section.

C. Comparison between Δ and p

It is instructive to consider the results that would be yielded if Δ were applied to a test composed of uniformly difficult items, administered with a very stringent time-limit. By hypothesis, the items are all equally difficult. But the values of Δ will be considerably higher for the later items of the subtest—because, in a test administered with a stringent time-limit, the values of M_t for those attempting the later items will considerably exceed the values of M_t for those attempting only the early items. In this situation, then, or wherever the value of N_t for later items is much smaller than Base N because of the speed factor, the use of Δ is not appropriate; the simpler measure, p ($=100\ N_c/N_t$) is likely to yield values much more nearly correct. In most tests, of course, the discrepancy between N_t and Base N is generally due to both the limitation in time and the increase in inherent difficulty of the items. Whether Δ or p serves better in such cases would appear to depend on the degree to which "speed" or "power" is determining the discrepancy between Base N and N_t.

It may be suggested once again that in a "power" test, the time-limit is likely to be ample; so that (if the examinee follows the directions to mark what he considers the best choice for *each* item) N_t is likely to approach Base N—in which case both Δ and p will yield similar results. In other words, p is definitely superior to Δ in the case of a pure-speed test; but Δ is not likely to enjoy an equally great advantage in the case of a pure-power test.

Continuing the comparison between Δ and p, it appears that the chief shortcoming of p is its complete neglect of ability-changes in the sample on which it is based; except in a pure-speed test, some adjustment of p for the value of M_t in the sample attempting the item would appear desirable. Another defect of p is that it is not expressed in terms of equal units; thus, the difference between two p's of 90 and 95 is really greater than the difference between two p's of 50 and 55. The force of this objection is somewhat weakened, however, in view of the fact that most values of p lie within a more or less restricted range. Advantages of p include the fact that (*a*) it is non-technical and readily understood; (*b*) its *PE* is known and easily calculated; and (*c*) it serves with markedly greater validity than Δ in the case of tests which place considerable emphasis upon speed. Unfortunately, neither p nor Δ can be trusted to yield entirely valid results in all instances.

Probably the best way to determine the difficulty of an item is to make use of a suitable experimental technique. One method which has proved effective in this Project's experience is to use an extremely liberal time-allowance on an experimental form of the test. This method, however, is not applicable in the case of a test which places a premium on speed of performance, since an ample time-allowance would permit many individuals to review and correct their answers—which is not possible under the regular time-allowance for such a test. Another possible objection to the use of a single form of the test with liberal time-allowance is that some individuals tend to become discouraged by repeated encounters with increasingly difficult items; as a result, such individuals either fail to attempt the later items, or fail to attempt them with normal effort and zeal. Under favorable conditions of administration and rapport, however, this problem does not seem to be seri-

ous.—A theoretically more rigorous, but practically more troublesome, procedure is to employ at least two arrangements of the items of the experimental test. Items appearing near the end of the test in one arrangement may be placed in a more advantageous position in the other arrangement. If the data for an item in its different positions agree, presumably the results are free from position-effect; if not, the data based on the item in its earlier position would generally be accepted.

All the procedures outlined above require that the experimental test be administered in a typical or fair sample (typical both as to average level of ability and as to range of ability). Both of the procedures aim to determine the difficulty of the item when attempted by all members of the sample. If it is desired to know only the difficulty of the item in its final position in the test under the regular time-limit, then a routine administration of the test to a fair sample is, of course, all that is needed.

If p and Δ are both computed for the items of a subtest, a strong correlation between p and Δ will ordinarily be observed; and it becomes tempting to infer that p and Δ are—for practical purposes—equivalent. This inference is correct, provided that the values of M_t and σ_t for the various test-items differ only slightly, and provided that p remains within a fairly restricted range around 50 (roughly, within the range 25-75). The formula for Δ, it will be recalled, is $\Delta = M_t + x'\sigma_t$. For any fixed values of M_t and σ_t, use of this formula shows that the difference in Δ for two items with p's of 85 and 95 is about 2.5 times the difference in Δ for items with p's of 45 and 55; throughout the range of difficulty, there is a point-to-point correspondence between p and Δ, but the correspondence is not linear. The point-to-point correspondence between p and Δ is not seriously disturbed by such differences in σ_t as generally occur from one item to another; but the effect of differences in M_t may be considerable. This is illustrated by the following data (based on a national sample of 500) for three items from the Sentence Completion Test of the General Classification Test (Form I):

Item No.	N_t	M_t	p	Δ
9	499	13.05	49	13.2
21	429	13.80	49	13.9
30	208	15.01	49	15.1

Although p is constant, the values of Δ are definitely not constant—being 13.2, 13.9, and 15.1, respectively. The maximum difference between the Δ's is 1.9. In general, a difference between Δ's of 1.00 or more respresents a difficulty-difference which is both subjectively perceptible and practically significant. Thus, it is not generally justifiable to adopt the easy view that the choice between Δ and p is of no consequence, on the ground that the two measures yield equivalent results. The policy of this Project has been always to report p; sometimes Δ has also been given. Both measures are useful; largely because p is less technical and more readily comprehensible, it has received some preference from this Project.

D. Biserial Correlation ($r_{bis.}$) between Item and Criterion

In practice, the most important single product of item analysis is the correlation between each item and the criterion. Usually the criterion is the score on the subtest of which the item is a part. In the work of this Project, the item-criterion correlation is measured by biserial r. Biserial r is computed for each item, except items for which p exceeds 95 or falls below 5 per cent. (The reason for excluding items with very high or very low values of p is that the biserial r for

such items is subject to excessive fluctuation of sampling—see Table 6.) The formula employed for the computation of biserial r is

$$r_{bis.} = \frac{M_c - M_t}{\sigma_t} \cdot \frac{p}{100(z)},$$

where

$r_{bis.}$, M_t, σ_t, and p are terms which have been previously defined;

M_c is the mean (transformed) criterion score of those who answered the item correctly; and

z is the ordinate of the unit-normal-curve at the point separating p (the percentage of successful attempts) from the remainder of the group attempting to answer the item.

The formula for biserial r given above is (except for modifications of notation) identical with the formula of J. W. Dunlap (*Psychometrika*, June, 1936, p. 51). As use of the subscript "t" in the formula above implies, the biserial r's computed by this Project are based on N_t (the number of cases attempting the item) rather than Base N.

1. *Some Statistical Aspects of Biserial* r

Biserial r is, in a sense, a measure of a hypothetical relationship: it measures the relation that would obtain between the item (X) and the criterion (Y) if the categorical pass-fail scores for the item were replaced by exact, quantitative scores distributed in a normal curve. It is assumed that the mean Y-value for the "pass" category falls on the hypothetical regression line (i.e., the regression of Y on the hypothetical values of X); similarly, it is assumed that the mean Y-value for the "fail" category also falls on this regression line; these two assumptions, together, are tantamount to the assumption of linear regression of Y on X. It may be observed that the assumptions with regard to linear regression and the normal distribution of X are not subject to empirical verification. Thus, while we may probably be right in the application of biserial r, we cannot be sure.

It should be noticed that the "standard error of estimate" (the S.D. of the array of Y-values for a given category of X) is not generally the same for the "pass" and the "fail" categories—unless $p = 50$, or unless $r_{bis.}$ equals 0. Thus, while the hypothetical relation between X and Y may be characterized by homoscedasticity, the relation in the empirical chart giving values of Y for each category of X, is not. With an extreme dichotomy, the standard error of estimate of Y for the category of X containing the majority of cases becomes quite large.

In computing an individual's score, the difference between passing and failing an item is represented by a uniform value; viz., the difference between 1 and 0. But in the calculation of biserial r, the value assigned to a pass or a fail is based on the normal curve, and the difference between these values is not uniform from one item to another. Thus there is an inconsistency between the practice in determining the individual's test-score and the practice in computing biserial r. The inconsistency is probably not practically important, and it may well be that both practices are justified; the logical contradiction, nevertheless, remains.

Another logical contradiction arises from the fact that all items are counted for all cases when determining each individual's total subtest-score; but only the sample represented by N_t is used in calculating the correlation between the item and subtest-scores. In the latter in-

TABLE 5

VALUES OF BISERIAL r FOR TOTAL SAMPLE VS. SAMPLE REPRESENTED BY N_t, TOGETHER WITH RELATED STATISTICS

Test	Item-Position	$r_{bis.}$ (Based on Total Sample)	$r_{bis.}$ (Based on N_t)	M_t	σ_t	N_t	Base N (Total Sample)
Sent. Completion	Middle*	.32	.30	13.2	3.92	491	500
Sent. Completion	Three-quarter*	.51	.37	14.2	3.66	369	500
Sent. Completion	Final*	.67	.58	15.0	3.78	208	500
Opposites	Middle	.65	.64	13.1	3.91	494	500
Opposites	Three-quarter	.42	.38	13.5	3.73	459	500
Opposites	Final	.33	.23	14.0	3.68	364	500
Analogies	Middle	.45	.45	13.0	3.90	499	500
Analogies	Three-quarter	.22	.20	13.1	3.89	491	500
Analogies	Final	.44	.42	13.3	3.80	443	500
Reading	Middle	.58	.56	13.0	4.00	498	500
Reading	Three-quarter	.31	.28	13.3	3.98	449	500
Reading	Final	.42	.39	13.5	4.14	310	500
Arith. Reasoning	Middle	.66	.65	13.1	3.97	493	500
Arith. Reasoning	Three-quarter	.67	.60	13.6	3.90	415	500
Arith. Reasoning	Final	.42	.39	13.8	4.24	220	500
Block Counting	Middle	.77	.65	13.9	3.59	429	500
Block Counting	Three-quarter	.89	.66	16.0	3.20	232	500
Block Counting	Final	.72	.52	16.7	3.41	106	500
Mech. Comprehen.	Middle	.30	.29	13.0	3.95	498	500
Mech. Comprehen.	Three-quarter	.44	.29	13.5	3.78	452	500
Mech. Comprehen.	Final	.48	.34	14.1	3.81	338	500
Surface Develop.	Middle	.77	.59	14.6	3.07	378	500
Surface Develop.	Three-quarter	.96	.81	16.0	2.61	251	500
Surface Develop.	Final	.68	.63	16.4	3.17	95	500

* The "middle" item is midway between the first and last item of a test or subtest; thus, for Sentence Completion (consisting of 30 items), the middle item is taken as item no. 15. Similarly, the "three-quarter" item is three-fourths between the first and the last item. The "final" item, of course, is the last item of a test or subtest.

stance, the individual's failure to attempt an item results in his exclusion from the sample, when the unattempted item is under consideration; in the former instance, the individual's failure to attempt the item is counted the same as an explicitly incorrect answer. Again, both practices may be justifiable, but the logical contradiction seems to be worth observing.

2. *Effect of Use of* N_t *vs. Base* N *in Formula for Biserial* r

We mentioned above that the biserial r's computed by this Project are based on N_t (the number of cases attempting the item) rather than Base N. The arguments for the use of N_t in preference to Base N are much the same as were presented in connection with the calculation of p, and will not be repeated here. It is of interest to observe that the use of N_t results in values of biserial r which are, in general, *lower* than would be obtained by the use of Base N; hence the use of N_t may be said to yield comparatively "conservative" values of $r_{bis.}$. Table 5 illustrates this fact for items from eight tests or subtests of the Navy Basic Classification Test Bat-

tery, Form I. The data in Table 5 are based on a national sample of 500 cases, drawn from six naval training stations.

A practical disadvantage of the use of N_t instead of Base N is that M_t and σ_t must, in general, be calculated separately for each item (since N_t is, in general, different for each item). If Base N were employed, M_t and σ_t in the formula for biserial r could be replaced by a single mean and standard deviation for the total sample. A minor disadvantage of the use of N_t is the (usually slight) increase in the "probable error" of $r_{bis.}$. The probable error tends generally to be somewhat larger, first, because N_t is generally smaller than Base N; and second, because $r_{bis.}$ itself is generally somewhat smaller when based on N_t instead of Base N. (The formula for the probable error of $r_{bis.}$ is given in Section 5,b below.) Consistency in the system of computations suggests that if p is based on the sample represented by N_t, then biserial r should also be based on the same sample.

3. *Meaning of Biserial* r *in Terms of "Internal Consistency" and Item-Validity*

Unless an external criterion is employed, the biserial r for an item is the biserial correlation between the item and a test-score—usually the score on the subtest of which the item is a part. Since the subtest score is simply the sum of the scores on the individual items, it is apparent intuitively (and can be proved statistically) that the correlation between item and subtest is an outcome of the correlations between the item and each of the other items of the subtest.[4] In other words, the item-subtest correlation serves as a measure of the functional consistency between a given item and the other items of the subtest. If the item-subtest correlation (biserial r) for a particular item is high, then that item is highly consistent or "homogeneous" with the other items of the subtest. If the item-subtest correlations of all the items are high, then all the items are highly consistent with each other, and the "internal consistency" or homogeneity of the entire subtest is high. High internal consistency of a subtest results in a high "split-half" reliability coefficient for the subtest. Theoretically, the length of a subtest affects the measure of internal consistency or homogeneity of an item; the influence of this factor will be evaluated in section 5,d below.

While the internal consistency or homogeneity of an item is of interest and importance, the *validity* of the item is of still greater consequence. For the purposes of this memorandum, item-validity refers to the correlation between an item and an *external criterion* (i.e., a measure of practical performance, as distinguished from a test-score). Ordinarily, we lack any direct measure of the validity of an individual test-item; what we usually have is only the biserial r between the item and its subtest. To the degree, however, that the subtest is valid (i.e., correlates with the external criterion), it is likely that the item-subtest correlation provides an indirect indication of the degree of validity of the item. When the item-subtest and item-validity coefficients are both available, it is found that the items with high subtest-correlations are also, in general, those which correlate well with the external criterion (see this Project's Memorandum No. 12); however, the correlation between item and external criterion is usually appre-

[4] Theoretically, the "weight" or standard deviation of each item also enters in determining the value of biserial r, but the error which results from neglect of these weights is practically negligible.

ciably lower than between item and subtest. The higher the correlation between subtest and external criterion, the more confidence may be placed in the item-subtest correlation as an indication of item-validity. The limitations of the item-subtest correlation (biserial r) as a measure of item-validity are discussed at some length in section 5,g below.

4. *Choice of Test-Criterion*

As already mentioned, the usual practice is to correlate each item with a test-score rather than with an external criterion. If a test is composed of several subtests, and especially if only a single total score for the test is recorded, the question arises whether the test-criterion for an item should be the *subtest* of which the item is a part, or the *total* test. This question is of some practical importance, since many of the test-scores in use in the Navy are total scores based on two or more subtests.

In favor of correlating each item with *total* test score is the argument that, unless this is done, there is no guarantee that the total score will represent a self-consistent, unitary ability. This argument however, seems to us to put the cart before the horse. We do not normally combine w, x, and y into z, and then insist that z should be made homogeneous; we first determine whether w, x, and y tend to form a homogeneous set, and if they do, we may then prefer to combine w, x, and y into a single total score.

Let us suppose, however, that w, x, and y *are* fairly homogeneous, and *have* been combined into a single score, z. Should each item for subtest w be correlated against the total score, z—or against the score on subtest w—or against z and also against w? The answer, we presume, depends on the degree of correlation between w and the remaining tests of the set. Unless this correlation is very high, we should assume that w probably measures some aspect or aspects of a practical, external criterion, which are not equally well measured by x or y. If so, it would appear desirable to maintain (and emphasize) the independent or non-overlapping aspects of w, rather than to coalesce w with x and y. The unique or independent contribution of w can be better preserved by correlating each item in w against the score in subtest w—rather than against the score in a composite or total test, z. This reasoning would appear to justify the policy of correlating each item against the score on the appropriate subtest—provided, of course, that the total test is, or can be, divided into subtests, and that the scores on each subtest are sufficiently reliable. In the absence of reliable subtest-scores, one is faced with three alternatives: (*a*) correlating each item against scores on the total test or a combination of subtests—this has the disadvantages noted above; (*b*) correlating each item against the unreliable subtest scores—here question arises whether the results are worth the labor involved; and finally (*c*) omitting analysis of items in the unreliable subtest. The last-named alternative should be adopted only as a temporary expedient, pending the development of a subtest which should be adequately reliable.

The analysis above has assumed that the practical, external criterion which the test or subtest aims to measure is complex, rather than purely unitary or self-consistent. There does not seem to us any reasonable doubt that a practical criterion is generally complex, and analyzable into several distinct components or sub-criteria.

5. *Factors Affecting the Interpretation of Biserial* r

Listed below are a variety of factors which may be considered with reference to their bearing on biserial *r:*

a. The percentage of successful attempts to answer the item (*p*).

b. The "probable error" or sampling fluctuation of biserial *r*.

c. The variability or "range of talent" of the group attempting the item.

d. Individual differences in speed of performance (number of items attempted).

e. The length of the test-criterion.

f. The reliability of the criterion.

g. The limitations of biserial *r* as a coefficient of item-validity.

A discussion of these seven factors follows immediately below.

a. *Biserial* r *in Relation to Percentage of Successful Attempts* (p)

1. The biserial *r* for a *very easy* item should be discounted for two reasons. First, the biserial *r* for a very easy item is subject to greater fluctuation of sampling than an equal *r* for an item of moderate difficulty (see Table 6 in section *b* below). Second, a given biserial *r* for a very easy item does not imply the same discriminative power for the item, as the same *r* for an item of moderate difficulty. Consider, for example, item no. 13 of the Sentence Completion Test of the GCT (Form I). In a national sample of 500 cases, the biserial *r* for this item was .71; $p = 95$. From the evidence of biserial *r*, this is an effectively discriminating item; but, since $p = 95$, the effectiveness of this item is limited to differentiating only 5 per cent of the group from the remaining 95 per cent. Superior discrimination by an item requires not only a high biserial *r*, but also a value of *p* not too far removed from 50.

TABLE 6

PROBABLE ERROR OF $r_{bis.}$
($N_t = 450$, $r_{bis.}$ and *p* as Specified in Margins of the Table)

Value of *p*	Value of $r_{bis.}$								
	.00	.10	.20	.30	.40	.50	.60	.70	.80
5 or 95	.067	.067	.066	.064	.062	.059	.056	.052	.047
10 or 90	.054	.054	.053	.051	.049	.046	.043	.039	.034
20 or 80	.045	.045	.044	.043	.040	.037	.034	.030	.025
30 or 70	.042	.042	.041	.039	.037	.034	.030	.026	.022
40 or 60	.040	.040	.039	.037	.035	.032	.029	.025	.020
50	.040	.040	.039	.037	.035	.032	.028	.024	.020

2. The biserial *r* for a *very difficult* item is subject to the same considerations as just presented for a very easy item. In the case of a very difficult item, however, some counterbalancing factors should be taken into account. Thus, (*a*) a large proportion of the responses to difficult items are likely to be guesses; to the extent that the guesses are correct, the item-criterion correlation is reduced. This reduction in item-criterion correlation does not imply that the difficult item is, by that much, an inferior or less well-constructed item; but rather that difficult items are subject to a special handicap. (*b*) A somewhat similar handicap derives from the fact that difficult items are likely to be placed toward the end of a test or subtest. The sample

that attempts these later items is often more selected or homogeneous (i.e., has a smaller σ_t) than the sample attempting the easier, early items; this tends to reduce the item-criterion correlation (see section *c* below). (*c*) A third consideration relates to the informational or experiential background required to answer an item correctly. It may well be that the background favorable for answering difficult items is less uniformly distributed among the sample than for easy items; this results in greater advantage to those who have had the favorable background for difficult items, thus leading (in an aptitude test) to a lower item-subtest correlation. Finally (*d*), the function or ability measured by a difficult item is likely to include factors not common to the remainder of the items of a subtest; for example, the difficult "opposites" items may place a heavy emphasis on knowledge of comparatively uncommon words; the difficult "analogies" may require information of a more specialized kind than the easy analogies; etc. This, of course, tends to reduce the correlation between the difficult item and the score on the subtest of which the item is a part.—If difficult items could be constructed free from these handicaps, it would be feasible to insist that difficult items should be revised or discarded if they fail to yield the same biserial r as good average or good easy items. But probably the factors mentioned above are well-nigh inseparable from difficulty.

Discretion is, of course, required in the acceptance of difficult items with low biserial *r*'s. It is just as true for a very difficult as for a very easy item that the item differentiates only a small part of the sample from the remainder. Moreover, some items are difficult and yield low biserial *r*'s, not because of factors inherently or necessarily associated with difficulty, but because of ambiguity, lack of adequate "distractors," incongruity with the remaining items of the subtest, etc.

b. *The "Probable Error"* (PE) *or Sampling Fluctuation of Biserial* r

Like any other statistical measure, the biserial r for an item is likely to fluctuate from one sample to another. Circumstances favorable to a small fluctuation or low "probable error" of biserial r include: (*a*) a large number of cases (N_t is high); (*b*) a value of $r_{bis.}$ which is itself high; and (*c*) a value of p (percentage of successful attempts) which is not too far from 50 (say between 20 and 80). Table 6 presents the *PE*'s for various values of $r_{bis.}$ from .00 to .80, when p varies from 5 to 95 per cent, and $N_t = 450$.

The value of N_t in Table 6 was set at 450, because this is a fairly typical value of N_t in the item-analyses conducted by this Project. Base N is usually 500, but the value of N_t for an item is likely to be smaller, because of omissions. In tests placing an emphasis on speed, the values of N_t for the items near the end of the test may be quite small (between 100 and 200); this of course increases very markedly the *PE's* of the values of $r_{bis.}$ for such items.

Perhaps the most noteworthy feature in Table 6 is the sharp rise in the *PE* of $r_{bis.}$ as p rises from 80 to 90 and from 90 to 95 (or, correspondingly, drops from 20 to 10 and from 10 to 5). Table 6 provides a concrete example of the high *PE*'s of biserial *r*'s for items having very high or very low values of p.

The formula for the *PE* of $r_{bis.}$ is —

$$PE_{r_{bis.}} = \frac{.6745}{\sqrt{N_t}}\left[\frac{\sqrt{p(100-p)}}{100z} - r_{bis.}^2\right],$$

where

N_t = the number of cases attempting to answer the item.
p = the per cent of successful attempts.
z = the ordinate of the unit normal curve

at the point separating p (the percentage of successful attempts) from the remainder of the group attempting to answer the item.

The formula given above may be found (with slightly different notation) in T. L. Kelley's *Statistical Method*, p. 249. Theoretically, the *true* value of p and of $r_{bis.}$ should be employed in the formula; since, however, the true values are not known with exactitude, we can do no better than to substitute the empirical values. The smaller the calculated *PE* of $r_{bis.}$, the smaller the error which the use of empirical values is likely to introduce.

Applications of the *PE* of $r_{bis.}$ assume that fluctuations of the value of $r_{bis.}$ follow a normal curve. The truth of this assumption has not been verified; but it seems likely that for large values of N_t and for values of $r_{bis.}$ which are not very high (not above .75), the assumption of normality does not entail excessive error.

In general, the main use of biserial r is to help decide whether an item should be kept *in* a test or *out* of a test. Suppose, for example, that one wishes to retain only those items for which the probability is small (say less than 10-in-100) that the observed $r_{bis.}$ was derived from a true $r_{bis.}$ below .35. If the true $r_{bis.}$ for an item is .35, then, if $N_t = 450$ and $p = 50$, the *PE* of the distribution of empirical values of $r_{bis.}$ derived from the true $r_{bis.}$ equals .0360. Referring to a table of the normal curve, it may be observed that only 10 times in 100 would empirical values at-or-above 1.90 *PE* arise from a true $r_{bis.}$ of .35. Hence, when $N_t = 450$ and $p = 50$, it is necessary to select items whose empirical values of $r_{bis.}$ equal at least $.35 + 1.9(.0360)$, or .418: in less than 10 times in 100 would empirical values of r_{bis} at-or-above .418 arise from a true value of $r_{bis.}$ below .35. This value of .418, obtained on the assumption that $p = 50$ and $N_t = 450$, may be contrasted with the value required when $p = 95$ (N_t still remaining 450). In this case, the *PE* of empirical values of $r_{bis.}$ (when true $r_{bis.} = .35$) equals .0633; and $.35 + 1.9\ (.0633) = .470$. It is thus necessary to have empirical values of $r_{bis.}$ equal to at least .470, in order to meet the requirement that less

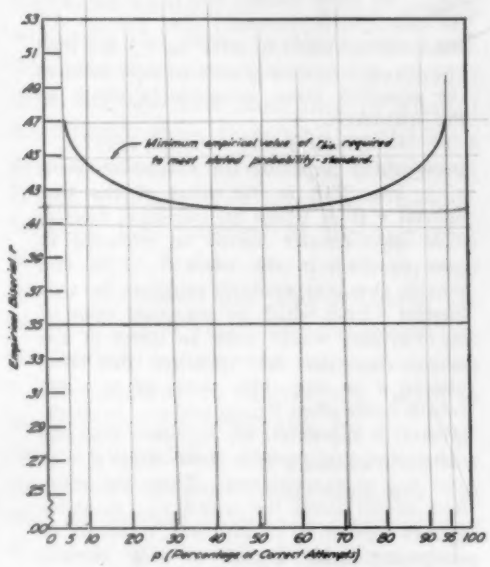

FIGURE 1. Minimum empirical value of $r_{bis.}$ required to assure that less than 10 times in 100 would the empirical $r_{bis.}$ arise from a true $r_{bis.}$ below .35, when $N_t = 450$, and p is as specified on the X-axis.

than 10 times in 100 would the accepted values of $r_{bis.}$ arise from a true $r_{bis.}$ below .35. In similar fashion, one may calculate the minimum empirical values of $r_{bis.}$ required, when p varies between 5 and 95 (N_t remaining 450). These minimum empirical values are plotted in Figure 1. Figure 1 shows clearly that, as p departs widely from 50, a definitely higher empirical $r_{bis.}$ is required to fulfill the stated probability-standard.

If one accepts all items for which the empirical value of biserial r is .35-or-greater, then for items with different values of p, one is really applying a *different* standard of true biserial r. This may be demonstrated as follows: Let the true biserial r from which an empirical value of

.35-or-greater would arise (say) 10 times in 100 be designated as r_∞. Then to determine r_∞, one solves the equation:

$$r_\infty + 1.9PE_{r_\infty} = .35,$$

where PE_{r_∞} is given by the same formula as cited above for $PE_{r_{bis.}}$ except that the symbol $r_\infty{}^2$ replaces $r_{bis.}{}^2$. In the formula for PE_{r_∞}, we shall assume that $N_t = 450$, and (in this specific instance) that $p = 50$. All the terms required to solve for r_∞ are thus known; and ordinary substitution reduces the equation given immediately above to the quadratic,

$$-.0604 r_\infty{}^2 + r_\infty - .27427 = 0.$$

from which (rejecting the extraneous root) $r_\infty = .279$. This is the value of the true biserial r from which an empirical biserial r of .35-or-greater would be expected to arise 10 times in 100, when $N_t = 450$ and $p = 50$. One may similarly calculate the true biserial r from which an empirical value of .35-or-greater would arise 10 times in 100 when $N_t = 450$ and $p = 95$; this true biserial r is .225. This value of r_∞, .225 (which holds when $p = 95$ and $r_{bis.} = .35$-or-greater) is approximately .05 lower than the value of r_∞, .279 (which holds when $p = 50$ and $r_{bis.} = .35$-or-greater). Thus, the selection of *all* items for which $r_{bis.} = .35$-or-greater results in a *different* standard of acceptance with regard to true biserial r—unless p remains constant. (For simplicity's sake, we have assumed that N_t remains constant at 450.) The curved line in Figure 2 shows the values of the true biserial r from which an observed $r_{bis.}$ of .35-or-greater would arise 10 times in 100, when $N_t = 450$, and p is as shown on the X-axis of the graph. It is clear from the graph that a different standard of $r_{bis.}$ must be applied to items with differing values of p, if a *uniform* standard of true biserial r is desired. Figure 1 gives the values of $r_{bis.}$ required to meet a uniform standard of true biserial r, on the assumption that the desired probability-standard is 10-in-100 or less, and that the minimum acceptable true biserial $r = .35$.

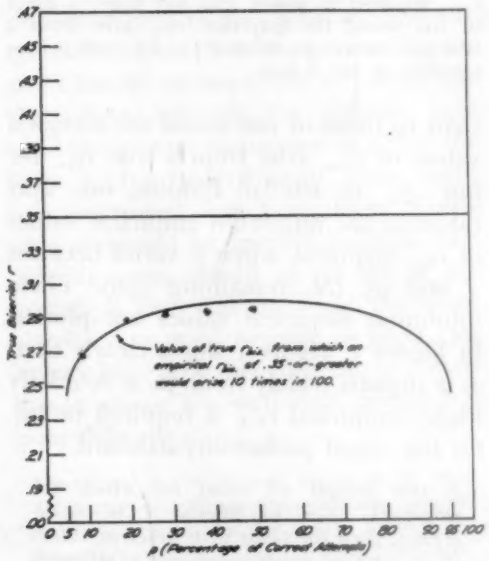

FIGURE 2. Value of true biserial r from which an observed $r_{bis.}$ of .35-or-greater would arise 10 times in 100, when $N_t = 450$, and p is as specified on the X-axis.

c. *Biserial r in Relation to Variability or "Range of Talent" of the Group Attempting the Item* (σ_t)

If the individuals of a group differ very widely in ability, it is comparatively easy to discriminate the better from the poorer; on the other hand, if the variability or "range of talent" in the group is narrow, discrimination becomes more difficult. The significance of this fact may be illustrated by data for two items from the Surface Development subtest of the Mechanical Aptitude Test, Form I (the data are based on a national sample of 500 cases). Both item no. 4 and item no. 20 of Surface Development have a biserial r of .59; but the variability of the group attempting item 4 (as measured by σ_t) is 3.76, while the variability of the group attempting item no. 20 is 3.07. If a correction were made for the limited variability of the group attempting item no. 20, the biserial r of .59 would rise to .67. Difficult items, appearing near the end of a test, are likely to be attempted by a more-or-less selected group having a low variability; this fact is one of those that should be considered when comparisons are made between the biserial r's of difficult versus easy items.

The formula used to make the correction given above is taken from T. L. Kelley's

Statistical Method, p. 225:

$$R_{xy}=\frac{r_{xy}\left(\frac{\Sigma_y}{\sigma_y}\right)}{\sqrt{1-r_{xy}^2+r_{xy}^2\left(\frac{\Sigma_y}{\sigma_y}\right)^2}}$$

In this formula, the capital letters refer to the statistical values in the more variable group; the subscripts x and y refer to item- and criterion-scores, respectively; in its present application, R and r refer to biserial (not product-moment) correlations. A graph giving the value of R, knowing r, σ_y, and Σ_y, has been published by H. A. Toops and H. A. Edgerton in the *Journal of Educational Research*, 1927, vol. 16, p. 382.

d. *Biserial* r *in Relation to Speed*

For tests which place a considerable premium upon speed of performance, two different types of item-homogeneity may be distinguished: first, homogeneity with regard to the type of function measured by the different items (apart from speed); and second, homogeneity with regard to the time required to answer each item.[5] In a group-testing situation, it is obviously not feasible to measure the time required by each person to answer each item; for this reason, the second type of homogeneity will not be considered in the present Report. A measure of homogeneity of the first type ("functional homogeneity") is ordinarily given by the value of biserial r for the item in question. Our purpose here is to examine the effect of the speed factor on the value of biserial r. The basic assumption in the discussion which follows is that an usuccessful attempt generally takes longer to make than a successful one; i.e., that wrong answers take more time than correct answers.

Suppose, now, that a given individual answers a certain item of a subtest incorrectly; not only does the individual lose credit on the item which he answered wrong, but he has comparatively less time in which to answer the remaining items of the subtest. Those persons who pass the item, on the other hand, not only obtain credit for this particular item, but also have more time to attempt later items; thus, those passing the item gain an advantage over those who failed. If we multiply this advantage several-fold (to take account of the fact that other items besides the particular one under discussion are also answered incorrectly), it is clear that the speed factor tends to *increase* the value of biserial r. This increase in biserial r is likely to to be especially noticeable for the later items of the subtest. Other factors besides item-position which determine the extent of increase in biserial r are (*a*) the degree to which speed determines scores on the subtest; and (*b*) the correlation between speed and "power" (i.e., ability to answer items at increasingly higher levels of difficulty—assuming that the later items of a subtest are progressively more difficult than the earlier).

By way of illustration, Table 7 compares the biserial r's for the first five vs. the last five items of (*a*) a non-speed subtest (Analogies, from the General Classification Test) and (*b*) a speeded subtest (Block Counting, from the Mechanical Aptitude Test); the data are based on a national sample of 500 cases drawn from six naval training stations. In Table 7 it is clear that the median biserial r for the last five items of Analogies is comparatively low (viz., .37—representing a drop of .12 from the median of .49 for the first five items); on the other hand, the median biserial r for the last five items of Block Count-

[5] Homogeneity with respect to speed may be defined similarly to homogeneity with respect to type of function; viz., by the correlation between the time required to answer the given item correctly and the time required to answer other items of the subtest correctly.

ing is comparatively high (viz., .63—a gain of .07 from the median of .56 for the first five items; this gain is made in spite of a drop in median σ_t from 4.0 to 3.4).

The conclusion of this section is that the item-subtest correlation ($r_{bis.}$) fails to yield a valid measure of the functional homogeneity of items in a test which places a premium upon speed. To eliminate the spurious influence of speed, a special administration of the test is necessary (see section VIII, C).

TABLE 7

BISERIAL *r*'S AND RELATED DATA FOR FIRST FIVE VS. LAST FIVE ITEMS OF THE ANALOGIES SUBTEST AND THE BLOCK-COUNTING SUBJECT

ANALOGIES						BLOCK COUNTING					
Item No.	$r_{bis.}$	N_t	p	M_t	σ_t	Item No.	$r_{bis.}$	N_t	p	M_t	σ_t
1	.47	500	92	13.0	3.9	I-A	.29	499	94	13.0	4.0
2	.52	500	38	13.0	3.9	I-B	.57	498	76	13.0	4.0
3	.11	500	83	13.0	3.9	I-C	.65	498	54	13.0	4.0
4	.61	500	45	13.0	3.9	I-D	.56	498	44	13.0	4.0
5	.49	500	85	13.0	3.9	I-E	.38	497	63	13.0	4.0
Median	.49	500	83	13.0	3.9	Median	.56	498	63	13.0	4.0
36	.30	469	62	13.3	3.9	II-U	.89	136	69	16.9	3.2
37	.54	463	26	13.3	3.8	II-V	.78	122	71	16.8	3.3
38	.30	457	59	13.3	3.8	II-W	.63	113	40	16.7	3.4
39	.37	452	47	13.3	3.8	II-X	.62	108	70	16.7	3.4
40	.42	443	27	13.3	3.8	II-Y	.52	106	76	16.7	3.4
Median	.37	457	47	13.3	3.8	Median	.63	113	70	16.7	3.4

e. *Biserial* r *in Relation to Length of Subtest*

Except when an external criterion is employed, the biserial r for an item is typically the biserial correlation between the item and the subtest of which the item is a part. The question of the present section is: To what extent is the biserial r for an item affected by the length of the subtest? We may judge the importance of this factor from an illustration. Suppose that subtest s (for "short") contains 30 items, and subtest l (for "long") contains 60 items; suppose further that the reliability coefficient of each test is .80; and that the average item-subtest correlation (biserial r) in each case is .50. If the number of items in subtest l were reduced from 60 items to 30 (the same as in subtest s), the item-subtest correlation of .50 would necessarily drop. But the drop would not be large: a calculation shows that the item-subtest correlation would fall from .50 to .46. (See J. P. Guilford, *Fundamental Statistics in Psychology and Education*, p. 287, formula 117.) Of course, if the original test l were not only twice as long as s but also of very low reliability, the effect of reducing l to the length of s would be more drastic. Since few of the tests in the Navy's aptitude testing program, however, have reliabilities below .80 (most reliabilities are considerably higher), it does not seem profitable to pursue this illustration any further. We conclude that, in the interpretation of biserial r, it is not necessary (except

under conditions not typical of Navy testing) to pay regard to the number of items in the test or subtest.

f. *Biserial* r *in Relation to Reliability of the Criterion*

The magnitude of the biserial r between item and criterion depends, of course, not only on the characteristics of the item itself, but also of the criterion. The question of the present section is: To what extent is the biserial r for an item affected by the reliability of the criterion? As in the previous section, we may judge the effect of this factor by an illustration: in this case, the figures will be based on two items from the Opposites and Mechanical Comprehension subtests of the Officer Qualification Test (Form 2) administered to a national sample of 561 cases. Both item no. 22 of the Opposites subtest and item no. 71 of the Mechanical Comprehension subtest have a biserial r of .50. The value of σ_t for each item is (to one decimal) 4.0. The reliability of the Opposites subtest is .91, of the Mechanical Comprehension subtest is .74. If we correct the raw biserial r's of .50 for unreliability of the criterion—a justifiable procedure, since an item should not be charged with random errors of measurement in the criterion—the corrected r_{bis}'s are, respectively, .52 and .58. Thus, the Mechanical Comprehension item now appears somewhat superior to the Opposites item. The change, in this case, seems fairly appreciable. In general, however, the effect of correcting for differences in reliability will usually be smaller, because most differences between reliability coefficients are smaller than in the illustration above. A comparatively minor factor, in this connection, is the magnitude of biserial r: the higher the biserial r, the greater the sensitivity to differences in reliability.

The statistically trained reader will recognize that the present section is, in a sense, a continuation of the previous section. In that section, we considered the effect, upon biserial r, of *reducing* the length (or reliability) of the subtest. In the present section, by correcting the item-subtest correlation for unreliability of the subtest, we in effect *increased* the length of the subtest to infinity. In both cases, the effect on the item-subtest r is brought about through a change in the reliability coefficient of the subtest.

The formula used to correct for unreliability or chance errors of measurement in the criterion is

$$r_{1c_\infty} = \frac{r_{1c}}{\sqrt{r_{cc}}}$$

where r_{1c} is the empirical biserial r between item and criterion, r_{cc} is the reliability coefficient of the criterion, and r_{1c_∞} is the corrected biserial r. This formula may be found in J. P. Guilford's *Fundamental Statistics in Psychology and Education*, p. 288.

g. *Limitations of Biserial* r *as a Measure of Item-Validity*

In this Report, the term "validity" refers to success in predicting individuals' scores or standing in a practical, external criterion which is itself valid. Ideally, each item should be correlated first against subtest score, to insure a homogeneous test; and second against a satisfactory external criterion, to insure validity. This Project's Memorandum No. 12 provides an example of such a double item-analysis. Unfortunately, the extra cost of double item-analysis, coupled with the difficulty and delay of obtaining measures on an assuredly valid external criterion, make it generally unfeasible to go beyond the usual item-subtest correlations.

As indicated in a previous section, the item-subtest correlation may ordinarily be accepted as an indirect indication of

validity, provided that the subtest itself is known to have a "satisfactory" correlation with a valid external criterion. Systematic, extensive data are required to support a quantitative definition of the term "satisfactory"; on the basis of this Project's Memorandum No. 11, perhaps a correlation of .45 or .50 may be suggested as a reasonable lower limit of "satisfactory." A correlation of .45 or .50, however, is by no means high enough to permit unqualified substitution of the subtest for the external criterion. Unless the correlation between the subtest and the external criterion is much higher than commonly observed, use of the item-subtest correlation (biserial r) as an indirect measure of validity raises the following questions:

1. What is the likelihood that an item with an acceptable or high biserial r (say above .40) would correlate low with a valid external criterion?

2. What is the likelihood that an item with a low biserial r (say below .35 or .40) would have a fair or high correlation with a valid external criterion?

The answer to the first question has already been given: it seems reasonably safe—in the usual case, where the subtest score is itself at least moderately valid—to consider a high biserial r as acceptable evidence of the external validity of the item. Theoretically, it is possible for an item to reproduce very faithfully that part or component of a subtest which is *un*correlated with the external criterion; in such a case, the biserial r between the item and subtest might be acceptably high (say .40-.50), yet the external validity of the item would be poor. But such a strong contrast between biserial r and external validity seems rather freakish. It could not possibly occur in more than a few cases (if it occurs at all); for if it did, the total subtest score (which is itself only a cumulation of item scores) could not correlate satisfactorily with the external criterion.

Similarly—in answer to the second question above—a low biserial r for an item will generally indicate a low external validity for the item, if the test itself has reasonable external validity (say a correlation of .45 or better). Three types of exception to this rule, however, suggest themselves:

1. An item may have genuine validity—in the sense that it measures well some important aspect of the external criterion—but be of a type formally unlike the other items of the test. In such a case, the biserial r for the item may be low, despite the fact that the item possesses a reasonable degree of true or external validity. For example, items 21-25 of the Reading Test, Form I, are based on a reading-paragraph which involves novelty and technical language; subjective analysis also suggests that the paragraph requires a certain degree of spatial ability. In these respects, this paragraph differs rather definitely from the remainder of the test. The low biserial r's for items 21-25 may be due, then, to their exceptional character, rather than to lack of external validity.

A common statement in mental testing is that the most efficient prediction is obtained when the elements of a battery correlate high with the criterion but low with each other. From this point of view, the occasional items which correlate satisfactorily with an external criterion but *low* with subtest score are ideal. Such items, however, are better made the basis for separate subtests,

rather than being retained in a group of other items with which they are not consistent. In this way the meaning of the total score in each subtest is left unblurred. A subtest which contains a heterogeneous mass of virtually uncorrelated items may be statistically efficient; but such a pot-pourri inevitably prevents clarity of understanding, hinders research, and delays progress.

2. An item may possess acceptable external validity, yet be of a difficulty-level such that the biserial correlation with subtest score tends to be rather low. The factors tending to depress the item-subtest correlation in the case of difficult items have already been considered in section *5a*, above.

3. An item may conceivably possess a validity so much greater than the other items of a subtest, that it correlates rather low with the other items, and hence also low with total subtest score. Unless the subtest itself is of very low validity, such an item would be in the nature of a freak—but a valuable freak. In the absence of direct correlation between each item and a satisfactory external criterion, subjective judgment should be on the alert for such items. The following two items from the O'Rourke General Classification Test (GCT) may possibly represent instances where the low bi-serial *r* is misleading:

> *GCT, Form B, Item 31:* John worked two times as long as Will and three times as long as Frank. What is the last letter of the name of the boy who worked the least
> J K L M N

> *GCT, Form C, Item 51:* LEAVES are to TREE as *Lungs* are to (1) heart (2) man (3) breathing (4) air (5) temperature.

The biserial *r*'s for these items are rather low (generally below .35[6]); yet to subjective judgment, both items appear very satisfactory. It is possible to object to item 31 on the ground that it is, formally, of a type occurring very seldom in the O'Rourke GCT; considerations of test-purity might therefore justify the exclusion of item 31. This objection does not, however, apply to item 51. It seems possible that item 51 correlates low with total-test score primarily because it calls for more careful, precise thinking than many of the other items of the test. If this is so, such exceptional merit should not be sacrificed to an indiscriminate, routine application of a minimum standard of biserial *r*. We should prefer to retain item 51, despite its low biserial correlation with total-test score. The final, objective justification for such a decision would consist, of course, in a satisfactorily high correlation between the item and a valid external criterion.

Biserial *r* has been discussed at length in this Report, because it is the most important single measure yielded by item analysis. We have been at some pains to present the instances in which the item-subtest correlation (biserial *r*) may fail to provide a good meaure of the validity of an item. It deserves emphasis, however, that if the subtest itself has a reasonable degree of external validity, the item-subtest correlation generally provides a useful measure not only of item-homogeneity, but also of item-validity. If this were not so, one of the main justifications for item analysis would be lost.

[6] See this Project's Report No. 5, pp. 4, 7. The median biserial *r* for the items of the O'Rourke GCT, Form B is .55, and for Form C is .58 (Report No. 5, pp. 54, 55).

V. INFORMATION CONCERNING THE ALTERNATIVES WITHIN EACH ITEM

THE INFORMATION concerning the alternatives within each item includes—

a. n, the number of cases choosing each alternative offered by an item; and

b. M, the mean (transformed) criterion-score of those selecting each alternative.

These two types of data may be discussed together, since the same considerations apply to each.

1. The first point to be emphasized is that n and M each provide extremely detailed information. The item-measures discussed up to the present—p, Δ, and $r_{bis.}$—are, so to speak, whole-item measures based on the entire sample (or that part of the sample represented by N_t). In n and M, however, we deal with the specific, individual alternatives within each item, and the sub-groups of persons selecting each particular alternative.

2. Because n represents the sub-group of N_t selecting a given alternative, it follows that n will frequently be fairly small. As a result, n, and also M (which is based on n) will frequently have high "probable errors"—i.e., to be subject to wide sampling-fluctuations. This emphasizes again the desirability of a large value of N_t, and, correspondingly, of Base N. The fact of sampling error should not be "conveniently ignored" when values of n or of M are under consideration.

The chief use of the M's for those choosing each of the alternatives within an item is in obtaining clues for item-revision (see section VII, B below). Since items for which biserial r equals .40 or greater are seldom revised, it is not customary, in the work of this Project, to calculate all the M's for such items. It is, however, necessary to compute M_c (the mean transformed criterion-score of those selecting the correct alternative) for every item, in order to calculate $r_{bis.}$ (see the formula for $r_{bis.}$ in section IV, D above). For items with values of $r_{bis.}$ below an arbitrary figure (say .40), M is computed for each group choosing each alternative (as well as for the group that "skipped" the item). The only exception to this rule occurs when n, the number of cases choosing an alternative, is less than 10. A mean based on fewer than 10 cases would obviously be too unstable to justify any serious consideration.

VI. NEED FOR INTERPRETATION

THE OBJECTIVE data of item analysis do not serve as a substitute for alert intelligence, but provide some facts for intelligence to work with. In best practice, each piece of item-information—whether this be $r_{bis.}$, p, Δ, n, M, M_t, σ_t, or N_t—is interpreted with due regard to the other pertinent data. An important fact to keep in mind when interpreting item-analysis data is the relative requirement which the subtest makes of "speed" vs. "power"; one indication of the "speed" requirement is the drop in N_t from the first item to the last item of the subtest. (In a test such as the Radio Code Speed of Response, where the response to each item must be made within a limited time, the number of "skipped" and omitted items gives an indication of the extent to which "speed" plays a part.) Another important consideration is the correlation between the subtest and a valid external criterion; only to the extent that this correlation is high may the item-subtest correlation ($r_{bis.}$) be safely taken as an index of item-validity.

The data of item analysis, while commendably objective, would be more complete if certain subjective information were also provided. For example, in undertaking to improve an item which has failed to yield a satisfactory biserial r, it would be useful to know *why* individuals selected (say) alternative no. 1 as their answer to the item, rather than alternative no. 2. Similarly, in improving the "distracters" within an item, better substitutes could probably be devised if we knew *what it is* about a certain alternative that attracts a disproportionate number of high-scoring individuals, while another alternative attracts a disproportionately small number of low-scoring. The data of item analysis do not supply direct information on such matters. In the absence of self-report from those actually taking the test, such points must be surmised from whatever resources are available of intuition, experience, training, and good judgment. Interpretation is required both with regard to the data which item analysis supplies, and the information which lies outside its realm.

VII. USES OF ITEM-ANALYSIS DATA

THE PREVIOUS sections have considered the characteristics and limitations of item-analysis data; it is time now to consider the uses of such data. The outline of uses in the present section is not, of course, intended to imply that item analysis supplies "all the answers," or that other techniques do not also make important contributions.

A. Provision of Objective, Quantitative Evidence Concerning Individual Items

Item analysis is the main source of quantitative, objective information about individual test-items. The objective nature of item-analysis data serves admirably in helping to settle arguments and objections concerning specific items —whether these arguments or objections are raised by experts, lay administrators, the examinees, or members of the public. Item-analysis data provide a convenient, practical basis for deciding which particular items are best suited for re-use in a subsequent form of a test; nor are the data wasted for rejected items, since frequently these items may be revised, with significant help from the data of item analysis (see section B below). The information yielded by item analysis is especially valuable (*a*) when the type of item employed is comparatively new and untested (e.g., certain "spatial ability" and "mechanical comprehension" items), (*b*) when the type of item employed tends to yield results not uniformly predicted by expert opinion (e.g., "analogies" and "mechanical comprehension" items), or (*c*) when a test has been constructed under adverse conditions (such as insufficient time for the most careful editing, lack of adequately trained and experienced personnel, etc.).

It is sometimes suggested or implied that the reliability coefficient of a test provides much the same information as item analysis. This is true, in the sense that the reliability coefficient of a test rather closely reflects the values of biserial *r* for the component items of the test. But the reliability coefficient fails to give any indication concerning differences in biserial *r* among the individual test items; and the information of item analysis is not restricted to the biserial correlation coefficient alone.

B. Improvement of Test Items

Item-analysis data are useful for locating weaknesses in a test-item, and for stimulating suggestions for improvement of the item. This can be illustrated by data for item no. 50 of the O'Rourke General Classification Test (GCT), Form C (a test formerly in use in the Navy). This particular item reads as follows:

> Diamonds are very valuable because (1) they are heavier than most jewels; (2) they are beautiful and rare; (3) they are cut and polished; (4) they refract light rays; (5) they are composed of pure carbon.

Data concerning this item were available from the naval training stations at Newport (Base $N = 239$), Great Lakes (Base $N = 210$), and San Diego (Base $N = 210$). In each case, the biserial *r*'s for this item were less than .20. The item-analysis data for this item, based on the subsample of recruits at the Newport Naval Training Station ("Newport NTS"), are given on the following page. We proceed to a detailed consideration of the item-data, with a view toward improving the item.

The item-analysis data of primary importance, from the point of view of improving a test-item, are the facts for *n* and *M*. Looking down the column headed "*n*" in the accompanying table, we find the number of cases who chose each alternative, respectively; looking

ITEM-ANALYSIS INFORMATION FOR A SPECIMEN ITEM

TEST: O'Rourke GCT FORM: C SAMPLE: Recruits at three NTS's
ITEM NO. 50 SUBSAMPLE: Recruits at Newport NTS

Alternative	n	M
o*	3	—
1	3	—
2	157	13.236
3	16	10.9
4	11	11.6
5	48	13.7

Base $N = 239$

$N_t = 238$
$M_t = 13.004$
$\sigma_t = 3.694$

$p = 66$
$\Delta = 11.5$

$r = .11$

* "o" means that the item was "skipped"; as indicated in the adjoining column headed "n," there were 3 cases who "skipped" this item.

ITEM NO. 50

Diamonds are very valuable because (1) they are heavier than most jewels; (2) they are beautiful and rare; (3) they are cut and polished; (4) they refract light rays; (5) they are composed of pure carbon.

down the column headed "M," we find the mean (transformed) criterion-score of those who chose each alternative. The number of cases—

a. who chose alternative (1) is 3. This is a quite negligible number. Evidently this first distracter failed to "draw," and should be replaced by another. The average criterion-score of those choosing alternative (1) has not been computed (this is in accordance with the rule that no M is computed unless n is 10 or greater).

b. who chose alternative (2) is 157. Alternative (2) is the correct answer, as indicated by the two lines within which the data for alternative (2) are enclosed. The average score (M) of the group choosing alternative (2) is 13.236; this is only slightly above the average score of the total group attempting the item ($M_t = 13.004$, as given in the list of data in the right-hand portion of the accompanying table. (The reason for calculating the value of M for alternative (2) to three decimals is that it is used directly in the computation of r_{bis}. This is not the case with the other M's.)

c. who chose alternative (3) is 16. The average criterion-score (M) of this group is 10.9—decidedly less than the average score of the total group attempting the item. This is a good distracter—though it would be better if it drew a larger number of cases at the same criterion-score level as the 16 who actually chose it.

d. who chose alternative (4) is 11. The average criterion-score (M) of this group is 11.6—again satisfactorily lower than M_t.

e. who chose alternative (5) is 48. The average criterion-score for this group is 13.7—which is *higher* than the average score of those who chose the *correct* answer. Evidently this alternative is more than merely ineffective: it is operating directly *against* a satisfactory biserial r for the item, since the individuals who choose this alternative make a higher score on the criterion than those who chose the correct alternative.

Examination of the data for item no. 50 leads to three suggestions for revision:

1. Replace the first alternative (which very few individuals chose) by a distracter possibly more attractive—e.g., by some such new alternative as: "They are ornamental."

2. Replace alternative (5) by a different choice, such as: "They are imported," or, "They are stylish."

3. Replace the word "valuable" in the body of the item by a term more clearly and unambiguously related to the correct answer; one possible suggestion is to replace "valuable" with "expensive."

With these suggested revisions, the item would read:

> Diamonds are very expensive because (1) they are ornamental; (2) they are beautiful and rare; (3) they are cut and polished; (4) they refract light rays; (5) they are imported.

The item, as revised, represents only the first step in the process of improvement. The next step is to actually try out the revised item. In the case of an item such as the one we have considered, the theoretical chances for improvement are greater than the chances for non-improvement or damage to the item: this item was so poor to begin with, that an intelligently executed revision should at least do no harm. The practical chances for improvement are aided by the fact that the revision was based on a substantial body of empirical information. Not all revisions, of course, have an equally good chance for success, nor can all revisions be expected to be successful at the first attempt. An item which fails to respond to treatment should be either drastically revised or eliminated.

On the basis of item-analysis results we can either (*a*) retain an item, (*b*) revise the item, retaining the item in the final test, (*c*) revise the item, but reserve it for further check before using it in the final test, or (*d*) eliminate the item. For immediate purposes, choices (*c*) and (*d*) are alike: the item is rejected or eliminated. In the following section, we shall compare the rejections made by item analysis with those made subjectively by expert judgment.

C. Item Analysis vs. Expert Judgment in the Elimination of Inferior Items

An interesting question is whether experts, confronted with a list of (say) 45 items of a pre-test, can select the 15 poorest items from the 45, with substantially the same results as yielded by item analysis. Two persons with considerable test-construction experience participated in the following experiment. Each was given a copy of Form X-2 of the General Classification Test, and asked to eliminate 12 items of the Opposites subtest and 15 items of the Analogies subtest (the number of items in the X-2 version of these subtests is 42 and 55, respectively). The directions to the experts were to—

> "use all *general* knowledge and any intuition that you may have, but not any specific memories as to how any particular item worked out in actual trial. The items which remain, after you have eliminated those which you consider least satisfactory, should be well-balanced in difficulty. Retain a few 'ice-breakers' which might otherwise be considered too easy. Your main criterion should be: 'Does this item measure well what we want the items of this subtest to measure?' Eliminate the items which meet this criterion least well."

As already indicated, the two subtests selected for critical examination by the two experts were the Opposites subtest and the Analogies subtest. The considerations leading to this choice of tests were as follows:

1. Under the time limits employed, neither of these tests was strongly affected by the factor of speed of performance;[1] thus, this possibly complicating factor is reduced to minor importance.

2. Both the persons available for the experiment were considered by others

[1] This statement is based on the following values of N_t: for the first item of the Opposites subtest, $N_t = 986$; for the last item, $N_t = 757$. For the Analogies subtest, the corresponding figures are 985 and 860.

to be reasonably expert in constructing and evaluating these two types of item.

3. The Opposites subtest is considered by the Project staff to be a fairly "predictable" test; that is to say, it is believed that item analysis rather seldom yields highly unexpected results for Opposites. The Analogies subtest, on the other hand, is considered a rather "unpredictable" test. We wished to see if the experimenter (H. C.) stated that he had no conscious memory of the item-analysis results at all; the other (E. H.) believed that she could recall some information about a few (perhaps three or four) of the items—but upon comparison (at the conclusion of the experiment) with the actual item-analysis results, it was found that some of E. H.'s supposed memories were either inaccurate or mistaken. For

TABLE 8

DISTRIBUTIONS OF VALUES OF $r_{bis.}$ FOR ITEMS REJECTED BY H. C., E. H., AND BY ITEM ANALYSIS

$r_{bis.}$ for rejected items	OPPOSITES			ANALOGIES		
	Items Rejected by			Items Rejected by		
	H. C.	E. H.	Item Anal.	H. C.	E. H.	Item Anal.
.80–.84						
.75–.79						
.70–.74						
.65–.69	3	1		1		
.60–.64	1	1		1	1	
.55–.59		1		1		
.50–.54				2	2	
.45–.49	3	4		4	3	1
.40–.44			1	1	4	
.35–.39			1	1	2	3
.30–.34	1	1	2	1		1
.25–.29	1	1	2	2	3	4
.20–.24						4
.15–.19			1			1
.10–.14						
.05–.09		1	2			
.00–.04	1	1	1	1		1
Below .00	1	1	1			
$r_{bis.}$ not computed*	1		1			

* $r_{bis.}$ not computed for items with values of p above .95. See text, section IV, D.

perimenters' rejections of opposites agreed much better with the results of item analysis than their rejections of analogies.

It will be recalled that the directions specified that the experimenters should not make use of specific memories of any actual item-analysis results for the items of the particular Opposites and Analogies subtests of this study. Neither of the experimenters had any substantial conscious memory of actual item-analysis results for these two tests. One experimenter, the item-analysis results for the two tests in this study were only part of a fairly steady flow of such results for various other tests. At no time, for either experimenter, had any special attention-value attached to the item-analysis results for the two tests. Finally, neither experimenter had seen the item-analysis results for the two tests of this study for nearly six months. In view of all these facts, it seems safe to say that the influence of memory, if any, could scarcely have been other than

quite weak and practically negligible.

The item-analysis results employed by the Project staff to eliminate the 12 least satisfactory Opposites and 15 least satisfactory Analogies were based on a sample of approximately 1,000 recruits, tested at the Naval Training Station at Sampson.

Table 8 presents a summary of the results of the experiment, in terms of the values of $r_{bis.}$ for the items rejected subjectively by H. C., subjectively by E. H., and with the aid of item analysis by the Project staff. It is obvious, from Table 8, that the items rejected on a subjective basis have, on the average, considerably higher values of $r_{bis.}$ than the items rejected with the aid of item-analysis results. Almost half the items rejected by the experts have biserial r's below .45; but practically all the items rejected by item analysis have biserial r's below this figure.

In favor of the experts it may be remarked that their selection of opposites and analogies was somewhat superior to chance. Of the 42 items in the Opposites test (Form X-2) and the 55 items in the Analogies test (Form X-2), the percentage of items with values of $r_{bis.}$ below .45 was 32.5 and 45, respectively. Of the experts' rejected items, a somewhat larger percentage had values of $r_{bis.}$ below .45; viz., about 40 per cent of the opposites, and 50 per cent of the analogies. According to these figures, the analogies hardly seem significantly less "predictable" than the opposites.

It seems fair to conclude from this experiment that, if the items of an experimental test or pre-test (such as Form X-2 of the GCT) already represent a selection by expert judgment, then further selection fails to duplicate satisfactorily the results obtainable with the aid of item-analysis data.

D. Improvement of Distribution of Item-Difficulty

By making a frequency distribution of the values of p or Δ for the items of each subtest, one can observe whether the distribution of item-difficulty suffers from skewness, gaps, or excessive concentration of items at any particular difficulty-level. If the experimental form of the test includes a sufficient range of item-difficulty and a sufficient surplus of items, imperfections in the distribution of item-difficulty can be corrected simply by the judicious elimination of selected items.[2]

E. Improvement of Reliability

Item analysis can improve the reliability of measurement by identifying the items of a subtest which are least homogeneous (have the lowest values of biserial r); other things being equal, it is these items which are most appropriately excluded from the final form of the subtest. The effectiveness of item analysis in improving reliability hinges on several factors:

1. The range of values of biserial r for the items of the original subtest: the greater the range, the greater the difference between the rejected and retained items, and the greater the possible rise in subtest-reliability.

2. The number of items in the original subtest: the greater the number of items, the more rigorous may be the standard for retention of items in the final form of the subtest.

3. The reliability coefficient of the original subtest: the higher the reliability

[2] In this connection, it is well to remember that the desired type of distribution of difficulty for the items of a subtest depends in part on the general level of item-subtest correlation (see this Project's Report No. 5, *Item Analysis of Navy Aptitude Tests*, pp. 52-53).

coefficient, the less likely there are to be many items with low values of biserial *r*, and hence the less likely that selection of items will improve reliability.

Factors *1*, *2*, and *3* are interdependent; thus, the more reliable the original test, the larger the number of items the original test must have if item analysis is to lead to an improved reliability coefficient.

Systematic quantitative data to illustrate the influence of the factors listed above are completely lacking. In order to ascertain the effect of item analysis in a practical case, we have calculated the reliability coefficients of two subtests (*a*) in an experimental version of the Navy General Classification Test (GCT) and (*b*) in a final version. The reliability coefficients of the subtests in the experimental version (Form X-2) are based on 300 recruits tested at the Naval Training Station at Sampson; the reliability coefficients of the subtests in the final version (Form 2) are based on 400 recruits tested at the Naval Training Stations at Farragut, Great Lakes, and Bainbridge. It is judged that these two samples of 300 and 400 recruits are comparable, though objective data to verify this view are not available. The two subtests chosen for study are the Opposites and the Analogies. These particular subtests were chosen because (*a*) in neither of these tests does speed of performance play a major role, under the time-limits employed (item analysis of the typical kind is best employed where the speed factor is small or *nil*); and (*b*) the Opposites subtest is frequently regarded as one whose items can be subjectively evaluated with fair success, while the Analogies subtest is regarded as one for which item analysis is especially desirable; it seemed desirable to observe results for these contrasting types of subtests.

A direct comparison between the reliability of the experimental and the final forms of the subtests would not be fair, because the experimental form of the Opposites subtest contains 42 items, while the final form contains only 30 items; similarly, the experimental form of the Analogies subtest contains 55 items, while the final form contains only 40 items. (In each case, the experimental form includes about 40 per cent more items than the final form.) Accordingly, we have estimated (by the Spearman-Brown formula) the reliability of the experimental forms *reduced to the same number of items as contained in the final forms*. The reliability of the reduced experimental form of the Opposites subtest is .854; of the final Opposites subtest, .908. The final Opposites subtest is thus about .05 more reliable than an experimental form of equal length. This difference is statistically significant at the 1 per cent level; it is also large enough to be of practical importance. In the case of the Analogies subtest, the reliability of the reduced experimental form is .847; of the final Analogies subtest, .880. The difference of .033 is not statistically significant at the 5 per cent level; nor, in our judgment, is it sufficiently large to justify the labor of item analysis. What may be termed an "insurance-factor," however, requires consideration at this point. If the items of a subtest are selected on a purely subjective basis, the choice of items may occasionally prove defective or "unlucky." This would result in an atypically low reliability coefficient for the final form of the subtest. Item analysis provides the information needed to prevent such an occasional downswing of reliability.

The results reported above show that item analysis is likely to lead to some improvement in the reliability of a test. But if the original (experimental) form of the test is already highly reliable, then a large surplus of items in the original test may be required, in order for item analysis to effect a significant increase of reliability in the final form. This conclusion is necessarily somewhat vague until such terms as "large surplus" and "highly reliable" can be quantitatively defined on the basis of extensive, systematic data. By the limited available evidence, cited above, a "large surplus" means a surplus greater than 40 per cent; and "highly reliable" means a reliability coefficient of about .90 or greater.

F. Improvement of Independence of a Test or Subtest

Not infrequently a test-battery will contain two tests or subtests which are designed to measure *different* abilities (or different aspects of some general ability), but which actually measure much the same functions. An illustration of this is found in the Mechanical Knowledge Test of the Basic Classification Test Battery. This test yields two scores, one of which is intended as a measure of electrical knowledge, the other, as a measure of mechanical knowledge. Actually, the two scores are rather highly correlated, indicating that the desired differentiation between electrical and mechanical knowledge has been only partially achieved (see this Project's Memorandum No. 13). Various methods may be tried to increase the independence of the two scores. One method, which may be dubbed "cross item-analysis," is to correlate each individual item of the test (*a*) with the Electrical score, and (*b*) with the Mechanical score. The items retained for the Electrical part of the test should, of course, correlate high with the Electrical score and low with the Mechanical; similarly, the items retained for the Mechanical part should correlate high with the Mechanical score and low with the Electrical. Since the differences in these correlations for a given item will probably not be very large the first time such an analysis is carried out (due to the impurity of the original Electrical and Mechanical scores), it is desirable that Base N (and, correspondingly, N_t) be large—well over 500—in order that the obtained differences may be reliably determined.

It should be added that the procedure outlined above has not yet, to our knowledge, been given any extensive trial. Actual application is required in order to indicate the degree to which the procedure may be useful.

G. Improvement of Correlation between Subtest and External Criterion

In the typical item-analysis, the biserial correlation is calculated between item and subtest-score; this serves to insure that the subtest will be composed of items which are homogeneous *inter se.* A second requirement, however, is that the subtest should be highly correlated with a valid external criterion. One possible means of improving the extent to which this second requirement is met is to correlate each item of the subtest with the external criterion, and reject all items for which the correlation with the external criterion falls below some set value. In this way, one should obtain a subtest which is, from the first analysis, satisfactorily homogeneous, and, from the second, as valid as is possible to obtain from the given assortment of items.

The procedure outlined above has, as yet, had only a very limited trial; accord-

ingly, it is not yet possible to estimate the degree of improvement which the method is likely to yield in actual practice. It should be recognized that the use of an external criterion typically involves many difficulties. Thus, the use of course-grades in service schools is subject to the difficulty that these grades are likely to vary in validity not only from school to school, but also from instructor to instructor. Furthermore, course-grades do not always reflect the proper balance between text-book knowledge and practical performance; and sometimes a heavy weighting of petty-officer qualifications in assigning final grades renders the grades of limited value as a criterion of what the test is intended to measure. A further important practical difficulty is that the use of an external criterion typically entails considerable delay, whereas scores on the test itself are immediately available.

H. Stimulation of Hypotheses and Insights

Item analysis yields results which, at least in some instances, are unforeseen and unexpected. If this were not so, there would be no point to item analysis. As it is, however, the unexpected continues to arise, and to furnish the stimulus for fresh hypotheses and insights. For illustration of this, we may refer to the discussion of the biserial r for difficult vs. easy items. So far as we know, the literature on aptitude testing does not warn that difficult items tend to be characterized by lower biserial r's than easy items. This result appears to be largely unexpected. In the attempt to explain this fact, a more explicit understanding is gained of the characteristics which make an item difficult: such as the complexity of mental functions involved, and/or the requirement of specialized knowledge. The question then arises whether the difficult items of a subtest, if segregated into a new subtest of their own, would show an improved biserial r with the new subtest scores; and concerning this question, hypotheses can presumably be offered both *pro* and *con*. Further study might lead to the conclusion that complicated mental functions are inherently less predictable than simple, or that the effect of specialized information on the biserial r of items has been over-rated, or to some other conclusion not yet envisaged.—Turning to another report of this Project (Report No. 5), we observe a marked superiority of the "Incorrect," as compared with the "Correct," items of a test on punctuation. Another finding in the same report is that the "proverbs" type of item is superior to all the other types in the O'Rourke General Classification Test (formerly in use in the Navy). These facts invite reflection and hypothesis-formation. As a further illustration, it is observed that certain analogy-items function much more efficiently than others; what are the causes for this difference? Finally, in the Mechanical Aptitude Test of the Navy's basic battery, it is found that items of the Block Counting and Surface Development subtests are characterized by high values of $r_{bis.}$, while the items of the Mechanical Comprehension subtest are characterized by low. Is this due to a "speed" factor? Does the "speed" factor alone explain the difference?—These illustrations indicate the fertility of item analysis in raising questions which, in turn, frequently lead to hypotheses and sometimes to insights. It goes without saying that the various hypotheses and supposed insights require verification by a broader collection of data or by specific experimental research.

VIII. RECOMMENDATIONS

A. Utilization of Item-Analysis Results

Item analysis provides a great deal of information, which cannot be properly interpreted and exploited without careful, thoughtful examination. The first recommendation of this section is that adequate time be allowed for the careful consideration and active utilization of item-analysis results.

B. Verification of Subjective Judgments concerning Items

When test items are being constructed and edited, various objections to specific items are likely to arise, and various points of excellence are likely to be remarked. Such objections or commendations are, of course, matters of judgment, which require verification by empirical data. Hence it is desirable that the various objections and approvals concerning a given item be *systematically recorded,* and later evaluated with the help of the objective item-analysis results. In this way, subjective hypotheses concerning what makes an item "good" or "bad" can be continuously checked, and a dependable set of judgmental criteria for the acceptance or rejection of items be built up.

C. Elimination of Effect of Speed upon Functional Homogeneity of Items

To the extent that a subtest places emphasis upon speed of performance, the item-subtest correlation ($r_{bis.}$) yields a spuriously high measure of the functional homogeneity of items in the subtest (see section IV,D,5). The correct value of r can be determined only by a special administration of the subtest, wherein all individuals are given the opportunity to attempt all items of the subtest. Three alternative procedures leading to this end are as follows:

1. Allow sufficient time for all (or practically all) the individuals in the sample to attempt each item.—The disadvantage of this method is that a goodly portion of the group will have time to review and check their answers to many items. In the final form of a "speed" test, such review and check would rarely be possible; hence it may be objectionable to allow such review in the preliminary or experimental form. If, however, only very few answers are actually changed, this objection may not be very important.

2. An alternative solution requires the use of an experimental test containing a large surplus of items. Suppose, for example, that the final form of the subtest is to contain (say) 50 items. One might ordinarily include a surplus of 50 per cent—making a total (in this instance) of 75 items. In the case of a speed test, however, it would be well to add to these 75 items an extra 50 or 75; these additional items should follow the first 75 of the test. The last 50 or 75 items would not be included in the item analysis. The criterion-score employed for the analysis of the first 75 items would be the total score on the first 75 items only. It is assumed that the extra items at the end of the test would keep everyone busy until time is called, thus preventing anyone from reviewing and checking his work. The time allowed should, however, be sufficient to permit everyone (or practically everyone) to attempt the first 75 items of the test.

A disadvantage of the procedure out-

lined above is that a large surplus of items must be prepared; these items serve no function other than providing "busy work" for those who answer most rapidly. Another disadvantage is that the answers to the experimental form of the test may require the space of an entire answer sheet; it is frequently convenient to have one answer sheet serve for several tests or subtests, instead of only one.

3. The third possible solution is really a variant of the one just described; the difference lies in the fact that *all* the items in the experimental form are analyzed. This is accomplished at the price of doubling the size of the sample taking the experimental form of the test. Suppose (as in the example above) that the final form of the subtest is to contain 50 items, and that the experimental form contains 75 items (numbered 1-75); to these are now added a *second* experimental form of 75 items (numbered 76-150). *One* sample is given items 1-150, in that order. The criterion-score for this group is the score on items 1-75; and only items 1-75 are analyzed. A *second* sample is given a form of the test in which first appear items 76-150, followed by items 1-75. The criterion-score for this group is the score on items 76-150; and only items 76-150 are analyzed. (It is assumed that all individuals will have time to attempt the *first* 75 items in each form of the test, but that none (or practically none) will have time to review and check their answers.)

Of the three methods proposed above, the first one is economical with respect to the number of items required, the amount of answer-sheet space needed, and the number of individuals who must be tested. If research shows this method to be adequate, it is the one which should generally be employed.

D. Time-Limits and Make-Up of Experimental Tests

When the experimental form of a test is to be subjected to item analysis, the following recommendations may be made concerning the time-limit and make-up of the experimental form:

1. The time-limit of the experimental form should be sufficient to permit all (or practically all) persons to attempt every item—except possibly in the case of a test intended to place a premium upon speed. (Administrative procedures suitable for a "speed" test have been described immediately above.) A small "pre-experimental" trial may be desirable in order to determine the proper time-limit for the experimental form.

2. The experimental form of a test should contain more items than the final form of the test, so that only the provedly best items need be retained in the final form of the test. The proper amount of surplus depends on various factors: (*a*) One factor is the degree to which the various characteristics of the items—especially ease (p) and the item-subtest correlation ($r_{bis.}$)—can be satisfactorily judged in advance: the lower the predictability, the larger the surplus required. Knowledge of the predictability of a set of items must, of course, be based upon previous experience with similar items. (*b*) A second factor is the amount of testing-time available. If it is necessary to interpolate the experimental form into an established testing-schedule, or if several experimental forms are to be tried out in a given sample, the number of surplus items may have to be kept at a minimum. (*c*) A third factor is the standard of excellence which the final test is expected to meet: the higher the standard of excellence, the larger the surplus required.—Quite clearly, it is im-

possible to lay down any fixed rule regarding the proper proportion of surplus items in the experimental form of a test; probably a surplus of about fifty per cent is the minimum that should be employed. Thus, if the final form of a test is to contain 50 items, the experimental form should contain at least 75. In this connection, it may be suggested that "there is safety in numbers"; that is to say, one is considerably safer with 50 extra items for a 100-item test, then one would be (say) with 5 extra items for a 10-item test. An unlucky original selection of items would far more often lead to an inadequate supply of good items in the second case than in the first.

3. Each item of the experimental form of the test should contain one or more extra "distracters" (incorrect alternatives or choices). Thus, if the final form of the test is to be composed of items containing five alternatives each, the experimental form might make use of items containing six alternatives each. In each item, the alternative which proves least effective could be excluded from the item as it appears in the final form of the test.

A possible objection to this recommendation is that the examinees may react differently to the remaining distracters, when the discarded distracter no longer appears in the item. Research is needed to determine the practical importance of this objection. Our judgment is that the use of additional distracters in the experimental form of a test is well worth an extensive trial.

4. The experimental form of the test should contain a somewhat larger proportion of very easy items than the final form, and a considerably larger proportion of difficult and very difficult items.[1] This precaution is desirable in order to provide adequate protection against a possibly unlucky selection of the few items at the extremes of difficulty. The special surplus of difficult and very difficult items is recommended, because difficult items seem especially prone to yield item-subtest correlations which are too low to be accepted (even when a rather lenient standard is applied). This is not, of course, uniformly true for difficult items in all tests; if results from previous experience are at hand, the proper proportion of difficult and very difficult items to include in the experimental form of the test may be adjusted in accordance with that experience. As a rough general rule, the proportion of very easy items in the experimental form should probably be about 1½-2 times as great in the final form, and the proportion of difficult or very difficult items about 2-3 times as great as in the final form.

[1] By "very easy" items is meant, in general, items with *p*-values of 85 or greater; by "difficult and very difficult" is meant items with *p*-values of about 30 or less. These figures apply to items of the usual multiple-choice type, with four or five alternatives in each item.

E. Size of Sample

In general, the size of the sample taking the experimental form of the test should be fairly large—never less than 500, and preferably larger.[2] A sample larger than the minimum of 500 is especially desirable if it is anticipated that the item-subtest correlations will be low (say around .30). With low biserial *r*'s, the error of sampling ("*PE*") of biserial *r* is larger: thus, it requires 612 cases to make a biserial *r* of .30 equally reliable as a biserial *r* of .45 based on

[2] The recommendation of a minimum of 500 cases is based on the group-testing situation, where it is less expensive to measure additional cases than it is to gamble on comparatively unreliable results. For an experimental test which must be administered to individuals one at a time (e.g., a performance test requiring accurate timing of numerous successive steps), practical considerations would force a reduction in the size of the sample employed.

500 cases (assuming that $p = 50$ in each instance). Another factor requiring consideration is that items with low biserial r's are likely to be revised. The process of revision brings into use the values of n and M for each alternative in each item (see section V); and reliable figures for n and M require a large value of N_t.

A sample larger than 500 is also desirable if the average value of $r_{bis.}$ for the items of the experimental test is high, but a still higher average value is demanded. The only way in which the still higher average value can be practically attained is to exclude not only items whose biserial r's are low, but also items whose biserial r's are moderately high. In this circumstance, it is essential that a fairly stable difference exist between the moderately high biserial r's which are excluded, and the ostensibly higher biserial r's which are retained—otherwise, when the test is applied in a fresh sample, the supposed increase in average $r_{bis.}$ will be found to represent nothing but a sampling fluctuation. To achieve stable, dependable differences in this situation calls for a large value of N_t (say a minimum of about 750).

F. Restriction of Item Analysis to Experimental Forms

The time-limit of the *final* form of a subtest is usually such that, if an item analysis is performed on this final form, the resulting values of p, $r_{bis.}$, etc. are questionable—especially for the later items of the test, which are the ones most affected by the speed-factor. It follows that item analysis is, in general, best restricted to the *experimental* form of the subtest.[3] The experimental form should be administrated with a sufficient time-allowance to permit all (or practically all) persons to attempt each item which enters into the analysis.

Two observations should be made concerning the values of biserial r for the items in the experimental form of the test vs. the items in the final form. First, the items in the final form of the test may differ from those in the experimental form, through the elimination of ineffective distracters. Second, although the final form of the test contains fewer items than the experimental form, the items in the final form are, by selection, more uniformly high in biserial r than the items in the experimental form; as a result, the score on the final form is likely to constitute a more reliable and generally superior criterion against which to correlate each item. Both these factors—viz., improved individual items, and an improved criterion-score—would tend to improve the biserial r of the item in the final form of the test. The values of biserial r from the experimental form of the test should, therefore, be generally interpreted as conservative estimates of the values that would be found in the final form of the test, if the final form were administered in such a way as to permit all (or practically all) individuals to attempt each item.

G. Discrimination in the Calculation of $r_{bis.}$

The usefulness of calculating biserial r for each item of a subtest depends in part on the homogeneity of the items constituting the subtest. In general, the reliability coefficient of the subtest serves as a useful index of such homogeneity. The reliability coefficient of a subtest should, accordingly, be taken

[3] If an item analysis, based on a large, fair sample, is available for the experimental form of the subtest, a subsequent item analysis of the final form is not likely to yield sufficient additional information to prove worthwhile. Since, however, the final form of the subtest does generally differ in some respects from the experimental form, a general over-all evaluation of the final form may well be desirable; this evaluation may be based on the shape of the distribution of subtest scores, the reliability coefficient of subtest scores, and the correlation of subtest scores with scores on other tests or on an external criterion.

into account, before an extensive program of item analysis is undertaken:

1. Calculation of the value of $r_{bis.}$ for each item is *not* likely to be especially useful, if the experimental form of the test is highly reliable (reliability coefficient = .90 or more). The reason for this is that such a test is, in general, already highly homogeneous, and can contain relatively few items which fall below an acceptable value of biserial *r*. The biserial *r* for each item of a highly reliable test may, however, be justifiably calculated, if an exceptionally high standard of excellence is required in the final form of the test; in such a case, it is essential (*a*) that the experimental test contain a large surplus of items, so that there will be a sufficient supply of items with very high values of $r_{bis.}$; and (*b*) that the size of the sample be unusually large, so that a dependable difference will generally exist between the biserial *r*'s of accepted vs. rejected items (see section E above).

2. Calculation of the value of $r_{bis.}$ for each item is of limited value if the experimental form of the test is highly unreliable (reliability coefficient below, say, .70). The purpose of calculating $r_{bis.}$ is generally to select a homogeneous set of items; but if the criterion itself is of questionable homogeneity (as is the case when the reliability of the criterion-score is low), then the usefulness of $r_{bis.}$ for improving homogeneity is correspondingly questionable. In this situation, the technique of "factor analysis" can be employed to identify such major clusters of homogeneous items as may exist.[4] A less thoroughgoing procedure would be to employ the item-subtest correlation ($r_{bis.}$) for each item as a tentative measure of homogeneity; and to supplement this by the correlation between each item and a homogeneous external criterion. The items which have the highest correlations with both the subtest score and the external criterion are probably those which are most nearly homogeneous.

3. It follows from the discussion in the two preceding paragraphs that the calculation of $r_{bis.}$ for each item is most likely to be useful when applied to experimental tests whose reliability coefficients are *moderately* high—say between .80 and .90. Such experimental tests should be subject to item analysis as a matter of fixed policy.

H. Determining the Reliability of the Experimental Form

An obvious implication of the discussion in sections F and G above is that the reliability coefficient of the experimental test should be known before item analysis is begun. The reliability of the final (shortened) form of the test may be estimated by the Spearman-Brown formula. In general, the final form of the test will have a reliability at least equal to the estimated reliability, if only because the speed-factor in the final form tends to increase the reliability coefficient.

[4] Since the score on each test-item is either "pass" or "fail," the correlations for a factor analysis would have to be based on fourfold tables. Such correlations generally have a high *PE*. Moreover, the correlations between individual items are, in general, found to be low; this also results in a high *PE*. It follows that, if a factor analysis is to be made, the number of cases measured should be considerably larger than for an ordinary item-analysis; perhaps a Base *N* of 1000 is the minimum that should be employed for dependable results. For tests requiring individual administration, the statistical requirements remain the same, but practical considerations would force the use of a much smaller sample.

I. Correlation with an External Criterion

The item-subtest correlation ($r_{bis.}$) should, whenever possible, be supplemented by correlating each item with a valid external criterion. This is especially desirable if the subtest itself has only a low correlation with the external criterion (say less than .45); because in such a case, there is danger of retaining items which, although homogeneous among themselves, are only slightly related to the external criterion which the test aims to measure. Similarly, it is desirable to correlate each item with a valid external criterion when the item-subtest correlations tend to be low (median $r_{bis.}$ below .45); because in this case, there is inadequate assurance from the item-subtest correlations that the items are sufficiently meritorious to be worth retaining. Knowledge of the correlations with the external criterion may also help to improve the homogeneity of the test (see section G,2 above).

IX. SUMMARY

THE PURPOSE of this Report is primarily to present a general, explanatory appraisal of the types of item-analysis data which have been supplied by Project N-106. A set of recommendations regarding item analysis is also presented.

A. Types of Information Supplied by Item Analysis

The information supplied in the item analyses of this Project may be classified into three main categories and various sub-categories, as follows:

1. Information concerning the item as a whole:

a. A measure of the *correlation* between the item and a criterion. The measure of correlation employed is the biserial correlation coefficient (symbolized as "biserial *r*" or "$r_{bis.}$"). The criterion employed is the score on the subtest of which the item is a part; if the test is not divided into subtests, then the score on the total test is employed. Occasionally, an external criterion (such as service-school grades) may be employed.

b. A measure of the *ease* of the item. This measure, symbolized by p, is defined as the per cent of successful attempts to answer the item. The formula for p is:

$$p = 100\ (N_c/N_t),$$

where N_c represents the number of individuals who answered the item correctly, and N_t represents the number who attempted to answer the item. The higher the value of p, the easier the item.

c. A measure of the *difficulty* of the item; the symbol for this measure is the Greek letter Δ (delta). In defining Δ, use is made of "transformed" criterion-scores; the essential features of these "transformed" scores are, first, that they correlate 1.00 with the original criterion-scores; second, that the mean of the total sample on the transformed scores is uniformly 13.0; and third, that the standard deviation of the total sample on the transformed scores is uniformly 4.0. Δ is expressed in terms of the same unit as the transformed criterion-scores, and is defined as that transformed criterion-score above which the percentage of cases equals p. The more difficult the item, the higher the value of Δ. The formula for Δ is given in section C,2 below.

d. The number of individuals who "skipped" the item. A person is judged to have skipped an item if he failed to record a response to the item, yet answered one or more subsequent items in the subtest (or in the total test, if the test is not divided into subtests). Normally, the number of cases skipping an item is small.

2. Information concerning the individual choices or alternatives offered by the item. This information includes:

a. The number of individuals selecting a given alternative in the item as the answer; this number is designated by the symbol, n.

b. The mean (transformed) criterion-score of those selecting a given alternative in the item; this mean is designated by the symbol, M.

3. Information concerning the sample attempting to answer each item. This includes:

a. N_t, the number of persons who attempted (or tried) to answer each item. An individual is considered to have "attempted" an item if he has recorded an answer either to this item or to any subsequent item in the subtest of which

the item is a part.[1] *All the item-data reported by this project* ($r_{bis.}$, p, Δ, M, n, M_t, and σ_t) *are based on the sample defined by* N_t. N_t is to be distinguished from "Base N," the total number of cases in the sample taking the subtest (or test).

b. M_t and σ_t, the mean and standard deviation, respectively, of transformed criterion-scores of those who attempted to answer the item.

The information yielded by the various measures defined above is more complete than is usually afforded by other item-analysis procedures. In our selection of measures, we have been guided by the extensive experience of the College Entrance Examination Board, under whose jurisdiction this Project operates. The characteristics of the various measures are summarized below.

B. Information concerning the Sample Attempting Each Item

1. *Number of Individuals Attempting Each Item* (N_t)

The number of persons attempting an item is symboloized by N_t. In a test placing a premium upon speed of performance, N_t diminishes rapidly from earlier to later items; this reduces the statistical reliability of the item-analysis data for the later items. The difference between Base N and N_t offers some indication of the degree of selection in the group attempting a given item; a much more direct and dependable measure of selection, however, is provided by M_t and by σ_t (see below).

2. *Mean* (M_t) *and Standard Deviation* σ_t *of Those Attempting Each Item*

The nature of the sample attempting each item is indicated directly by M_t and σ_t, the mean and standard deviation, respectively, of the transformed criterion-scores of those attempting each item. A group for which M_t exceeds 13.0 is superior to the total sample (Base N); a group for which σ_t is less than 4.0 is more homogeneous in ability than the total sample. In general, for the later items of a subtest, M_t tends to become progressively larger than 13.0, and σ_t progressively smaller. The factors which determine the trend in M_t and σ_t are:

1. The time limit for the test: the more sharply limited the time, the steeper the rise in M_t and the drop in σ_t.

2. The rate of increase in difficulty from early to later items in the subtest: the steeper the increase in difficulty, the greater the changes in M_t and σ_t. (This factor reflects the frequent unwillingness of examinees to guess on items which are quite beyond their ability.)

3. The correlation between number of items attempted (or speed of performance) and level of ability: the higher the correlation, the greater the changes in M_t and σ_t.

4. The homogeneity or internal consistency of the items in the subtest (as evidenced by the biserial correlation between the items and scores on the subtest): the higher the homogeneity, the greater the changes in M_t and σ_t.

Both M_t and σ_t are employed in the calculation of Δ; M_t is also of importance in the interpretation of p, and σ_t is pertinent in the interpretation of $r_{bis.}$.

C. Information concerning the Item-as-a-Whole

1. *Ease of Each Item* (p)

The chief question of interest in connection with p relates to the use of N_t

[1] If the test is not divided into subtests, the word "test" should be substituted for "subtest" in this definition.

versus Base N in the denominator of the formula, $p = 100\ (N_c/N_t)$. Neither the use of Base N nor N_t leads to uniformly satisfactory results. Base N is the proper denominator to use for p, if it is assumed that a person who fails to reach an item would also have failed to answer the item correctly had he been given time to attempt it. The use of N_t involves the assumption that, had more time been allowed, those who failed to reach the item would perform the same as those who did reach the item. The literature on the relation between "speed" and "power" gives better support to the assumption underlying the use of N_t than of Base N. If a test measures mainly speed of performance, the use of N_t is definitely preferable—since the use of Base N in such a case would result in p-values which reflect the position of the item in the subtest, far more than the inherent ease or difficulty of the item.

To the extent that N_t is smaller than Base N, the use of N_t in the formula for p results in a larger "probable error" of p. But the "probable error" of p is not ordinarily an important practical issue so long as N_t is fairly large—say 400 or more. For the later items of a test emphasizing speed of performance, N_t usually falls far below 400—unless Base N is unusually large (say 1,000 or more).

2. *Difficulty of Each Item in Terms of "Δ"*

Since the measure of item-ease (p) is not free from objection, a different measure, symbolized by the Greek letter "Δ," was devised by C. R. Brolyer and C. C. Brigham of the College Entrance Examination Board. The formula for Δ is: $\Delta = M_t + x'\sigma_t$. The terms M_t and σ_t in this formula have already been defined; x' is the unit-normal-curve abscissa corresponding to the value of p for the item (x' is positive for value of p below 50, negative for values of p above 50).

When the value of N_t for an item is fairly close to Base N (as is likely for the first half of the items of a subtest), both Δ and p yield substantially equivalent results. For the later items of a subtest, if N_t is considerably less than Base N, Δ is a better measure of item-difficulty than p, provided that the difference between N_t and Base N reflects mainly individual differences in "power" or level of ability; p is a better measure if the difference between N_t and Base N reflects mainly individual differences in speed of performance. Unfortunately, the relative influence of "power" vs. "speed" in determining the difference between N_t and Base N is not always definitely known. A solution to this problem is to administer an experimental form of the subtest in such a way that all (or practically all) individuals answer each item (see section IV, C).

Although a strong correlation usually appears between p and Δ, it is not recommended that the two measures be regarded as generally equivalent, especially if the values of M_t for the various items of the subtest differ considerably among themselves. Largely because p is the less technical and more readily comprehended measure, it has received some preference in the reports by this Project.

3. *Biserial Correlation* ($r_{bis.}$) *between Item and Criterion*

In practice, the most important unit of information yielded by item analysis is the correlation between each item and the criterion; in work of the present Project, this correlation is measured by biserial r ($r_{bis.}$). The criterion employed is usually the score on the subtest of

which the item is a part. Consistent with the practice in determining p, the item-criterion correlation ($r_{bis.}$) is based on N_t rather than Base N. The use of N_t results in values of biserial r which are, in general, *lower* than would be obtained by the use of Base N. Biserial r provides a measure of functional consistency between a given item and the other items of the subtest; such consistency has also been termed "internal consistency" and "homogeneity." If a test is divided into subtests, it is preferable to use the subtest-score as the criterion for the items in each subtest, rather than to employ total-test scores as a single, general criterion for all the items in the test.

Listed below are several factors which bear on the interpretation of the biserial r obtained for a test item:

1. The percentage of successful attempts to answer the item (p).—(a) If p is very low or high—say either below 10 or above 90—then the effectiveness of the item is limited by the fact that, at best, it can differentiate only a *small portion* of the sample from the remainder. (b) If the value of p for an item is very low (the item being very difficult), a low or moderate biserial r for the item sometimes deserves upward adjustment, because of certain technical and incidental handicaps which difficult items generally have to overcome (see pp. 18-19).

2. The "probable error" or sampling fluctuation of biserial r.—The statistical factors determining the magnitude of the *PE* of $r_{bis.}$ include the value of N_t, the value of p, and the value of $r_{bis.}$ itself. The *PE* of $r_{bis.}$ rises sharply as p rises above 80 or falls below 20. This is another limiting factor in the case of very easy or very difficult items. A common application of the *PE* of $r_{bis.}$ is in setting up a minimum acceptable value of $r_{bis.}$ for items which are to be retained in a test. A definitely higher standard must be set up when p is greatly different from 50 than when p is equal or nearly equal to 50.

3. The variability or "range of talent" of the group attempting the item.—The diminished range of talent of the sample attempting the later, more difficult items of a test tends generally to reduce somewhat the value of $r_{bis.}$ for such items.

4. The factor of speed of performance.—If speed of performance plays a considerable part in determining the score on a subtest, then the item-subtest correlation ($r_{bis.}$) tends to be spuriously high. A special administration of the test is necessary to eliminate the spurious influence of speed (see section VIII, C).

5. The length and reliability of the test-criterion.—Differences in the length and reliability of Navy tests are sufficiently small to render these factors practically unimportant.

6. The limitations of biserial r as a coefficient of item-validity.—The possibility was examined that (a) an item with an acceptable or high biserial r would have a low correlation with a valid external criterion; and (b) that an item with a low biserial r would have a fair or high correlation with a valid external criterion. Examples of these two possibilities have been given in the body of this Memorandum. As a general rule neither of the two possibilities is likely to materialize with significant frequency, provided that there is a satisfactorily high relation between external criterion and the test or subtest against which the items are correlated.

D. Information Concerning the Alternatives within Each Item

For the alternatives within an item,

the item-analysis data include n, the number of individuals choosing each alternative, and M, the mean (transformed) criterion-score of those choosing each alternative. Both n and M are subject to comparatively high "probable errors" or sampling fluctuations, since the group selecting any particular alternative is only a sub-sample of N_t. The values n and M give information useful in revising or improving test items.

E. Need for Interpretation

The objective data of item analysis do not serve as a substitute for alert intelligence, but provide some facts for intelligence to work with. In best practice, each datum from item analysis—whether this be $r_{bis.}$, p, Δ, n, M, M_t, σ_t, or N_t—is interpreted with due regard to the other pertinent data. For the improvement of test items, it is necessary to supplement the objective information from item analysis by shrewd judgments concerning the particular factors or qualities which make one item too easy and another too hard, or one "distracter" (incorrect alternative) excessively attractive to the superior individuals, while another is insufficiently attractive to the inferior. Such judgments are aided, but not supplied, by the data from item analysis. Interpretation is required both with regard to the data which item analysis supplies, and the information which lies beyond its scope.

F. Uses of Item-Analysis Data

The uses of item-analysis data may be briefly summarized as follows:

1. Item analysis supplies detailed, objective, quantitative information for each item. This information cannot be obtained by "expert judgment" nor by any manipulation of the reliability coefficient.

2. The objective, quantitative information from item analysis is well suited to help settle arguments or objections concerning specific items; and provides a convenient, practical basis for selecting items for subsequent forms of a test.

3. Item-analysis data provide information which is useful in revising and improving test items.

4. The distribution of item-difficulty can be improved with respect to symmetry, continuity, and average level, on the basis of the evidence provided by item analysis concerning the difficulty of each item.

5. The reliability of the test may frequently be improved by the judicious selection of items on the basis of item-analysis data.

6. The independence of the test from other tests in the battery may be improved by the application of a "cross item-analysis" technique (see section VII, F).

7. The external validity of the test can frequently be improved, if the item-analysis includes the correlation between each item and a valid external criterion.

8. The data of item analysis stimulate hypotheses and insights which are of use both in the construction of tests and the interpretation of test results.

G. Recommendations

1. Adequate time should be allowed for the careful examination and full exploitation of the information yielded by item analysis.

2. When test items are constructed, a systematic record should be kept of the supposed points of weakness and excellence of specific items. The data of item analysis should be employed to check on these subjective judgments.

3. One of the procedures described in the body of this Report (see section

VIII, C) should be followed in order to eliminate the spurious effect of speed of performance on the value of biserial *r*.

4. The following recommendations are made concerning the time-limits and make-up of experimental tests:

a. The time-limit of the experimental form of a test should be sufficient to permit all (or practically all) persons to attempt every item. Modifications of this rule, for tests which are intended to place emphasis on speed of performance, are considered in section VIII, C.

b. The experimental form of a test should contain more items than the final form of the test, so that only the provedly best items need be retained in the final form. In general, a surplus of at least fifty per cent is desirable. This surplus should be considerably larger for the very easy, and for the difficult or very difficult items of the experimental test (see section VIII, D).

c. Each item of the experimental form of the test should contain one or more extra "distracters" (incorrect alternatives or choices).

5. The size of the sample taking the experimental form of a group-test should be large—never less than 500, and preferably larger.

6. Item analysis is, in general, best restricted to the experimental form of a test or subtest, rather than the final form.

7. Full item analysis should not generally be applied to tests which, by the evidence of a high reliability coefficient (over .90), are already highly homogeneous. Item analysis is most likely to be useful when applied to experimental tests whose reliability coefficients are moderately high —say between .80 and .90. An obvious implication is that the reliability coefficient of the experimental test should be known before item analysis is begun.

8. The item-subtest correlation ($r_{bis.}$) should, whenever possible, be supplemented by correlating each item with a valid external criterion. This is especially desirable if the subtest itself has only a low correlation with the external criterion (say less than .45), or if the item-subtest correlations tend to be low (median $r_{bis.}$ below about .45).

APPENDIX

Specimen Item-Analysis Sheet

On the following page is given the item-analysis sheet for item no. 96 of the Mechanical Knowledge Test, Form I. Some of the entries on the item-analysis sheet (namely, "Card Number," "Date Tabulated," and "Operator Number") are for office use only, and need not concern us. On the item sheet, the item number is recorded as $\begin{matrix} 0 \\ 9 \\ 6 \end{matrix}$ (meaning "96") in a box in the upper-left corner.

The item-analysis sheet includes a few more details than are given on page 29 for item no. 50 of the O'Rourke GCT (Form C); in particular, one observes columns headed "Σx" and "Σx^2," together with figures at the foot of these columns. The column headed "Σx" gives the sum of transformed scores[1] of those selecting the alternative indicated under "Code." Thus, the sum of the transformed scores for the 3 cases who "skipped" this item (Code "O") is 16; the sum of the transformed scores of the 288 cases who chose alternative no. 1 is 4215; etc. The column headed "Mean" gives the mean transformed score of those selecting each alternative; this is obtained simply by dividing the value of "Σx" by the value of n in the column adjoining. The column headed "Σx^2" gives the sum of squares of the transformed scores for the individuals selecting the alternatives indicated under "Code." Thus, for the 3 cases who "skipped" this item (Code "O"), the sum of squared transformed scores is 88; for the 288 cases who chose alternative no. 1, the sum of squared transformed scores is 65683; etc.

The figures at the foot of the columns headed "n," "Σx," and "Σx^2" are sums of the figures in the respective columns. The sum of the n-column gives the total number of cases attempting the item, or N_t; in this particular instance, $N_t =$ Base N or 500. The sum of the Σx-column, when divided by N_t, gives the mean transformed score of those attempting the item; this mean, in the present instance, equals $6502 \div 500$, or 13.004 (recorded at the foot of the "Mean" column). Similarly, the sum of the Σx^2-column, 92570, is used in the calculation of σ_t, the standard deviation of transformed scores of those attempting the item.

The fact that alternative no. 1 is the correct answer for this item is indicated by enclosing the data for this alternative between *heavy lines*.

In the body of this Report, p represents a percentage; in the item-analysis sheet, however, p represents a proportion. It has seemed more convenient, in exposition, to use the percentage-form, although some formulas may be written somewhat more briefly by use of the proportion. We have also, in the body of this Memorandum, employed the symbol N_t (instead of n_t) to designate the number of cases attempting an item; and N_c (instead of n_+) to designate the number of cases answering the item correctly. The symbol "r_{bis}." for the biserial correlation coefficient is written on the item-analysis sheet as simply "r."

[1] "Transformed scores" are defined in section II, near the beginning of this Report.

ITEM ANALYSIS

Card Number 48

ITEM 0 9 6

TEST MECHKN

FORM I

BASE N 500

COLLEGE ENTRANCE EXAMINATION BOARD
Research and Statistical Laboratory
Princeton, New Jersey

Date Tabulated 6 44 Operator Number 9

Response	Code	n	Σx	Mean	Σx²
	0	3	16		88
	1	288	4215	14.635	65683
	2	71	848	11.9	10814
	3	45	458	10.2	4990
	4	93	965	10.4	10995

σ_t = 4.004

p = .58

$M_+ - M_t$ =

$\frac{M_+ - M_t}{\sigma_t}$ =

$\frac{p}{z}$ =

r = .60

△ = 12.2

$p = \frac{n_+}{n_t}$ $\sigma = \sqrt{\frac{\Sigma x^2}{n} - M^2}$

$r = \left(\frac{M_+ - M_t}{\sigma_t}\right)\left(\frac{p}{z}\right)$

96. In a gasoline engine, the gas mixture should explode in -- (a) the cylinder (b) the intake manifold. (c) the exhaust manifold. (d) the carburetor.

TOTAL TRIED (t)	500	6502	13.004	92570

Computed by K

Checked by

NOTES:—
1. x', based on p, is the distance from the mean along the baseline of the normal curve, in terms of unit standard deviation. If p is less than .5, x' is positive; if p is more than .5, x' is negative.
2. Compute means only for responses made by ten or more candidates.
3. Compute M_+ and M_t to three decimal places; means of all other responses, to one.
4. Record r to two places; △ to one.
5. Carry all other computations to three decimal places; i.e., $\frac{\Sigma x^2}{n}, \frac{\Sigma x^2}{n} - M^2, \sigma_t, p.$

42143

Psychological Monographs:

General and Applied

Characteristics and Uses of Item-Analysis Data

By

Herbert S. Conrad

Edited by Herbert S. Conrad

Published by The American Psychological Association

Psychological Monographs:
General and Applied

Manuscripts should be sent to the Editor. For suggestions and directions regarding the preparation of manuscripts, consult the following article: CONRAD, H. S. Preparation of manuscripts for publication as monographs. *J. Psychol.*, 1948, 26, 447-459.

Because of lack of space, the *Psychological Monographs* can print only the original or advanced contribution of the author. *Background and bibliographic materials must, in general, be totally excluded,* or kept to an irreducible minimum. Statistical tables should be used to present only the most important of the statistical data or evidence.

Correspondence concerning business matters (such as subscriptions and sales, change of address, author's fees, etc.) should be addressed to: DR. DAEL WOLFLE, American Psychological Association, 1515 Massachusetts Ave., N.W., Washington 5, D.C.

Published by The American Psychological Association